Federation Proceedings

VOLUME 11 *September 1952* NUMBER 3

JOINT SESSION OF THE FEDERATION

New York City, April 15, 1952

Chairman: D. B. DILL

TOPIC: WORLD FOOD PROBLEMS

FOOD AND POPULATION ILLUSTRATED BY OBSERVATIONS IN INDIA

W. H. FORBES

From the Harvard School of Public Health, Boston, Massachusetts

ALL of us who share in the activities of the six Societies have in common the desire of defining as exactly as we can, and mathematically if possible, the relationships between the various entities that interest us, whether these be molecules or whole organisms. The relationships I shall examine are those between large groups of men and the resources of their ranges. The understanding of the problems involved is vital to our civilization, but so far the amount of objective analysis and exact definition of the various factors has been small. This group is well qualified to think logically and constructively on this subject, and it is my hope to sharpen your interest in it.

The most important single concept in my discussion is that the ratio of food resources to the number of individuals determines the whole nature of a civilization. If this ratio, which I shall call 'food over population' (F/P), is high, the culture will have certain characteristics, e.g. individual longevity, social instability. If the ratio is low, the culture will have other and often opposite characteristics, e.g. poverty, short life expectancy and social stability. All that follows will be centered around this ratio, though I shall often discuss related variables rather than the ratio itself.

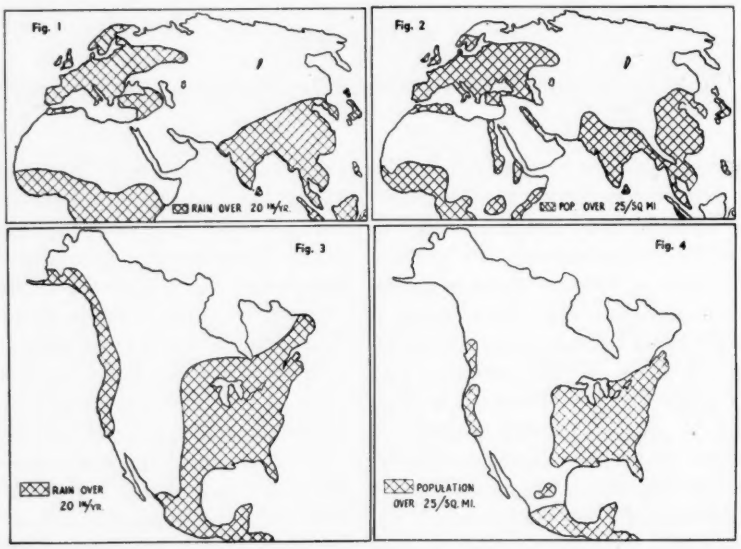
Food will be the main topic because at present and for the foreseeable future food is and will be the limiting resource. Oxygen is plentiful except upon the highest mountains, and water—that is water for drinking—is reasonably plentiful in all inhabited regions, with rare exceptions. Indeed in the hot dry regions a man will become very uncomfortable and even incapacitated in a few hours without water and may die in 24 hours under extreme conditions, so that he does not attempt to move in where there is no water to drink. People do not suffer from thirst when there is a drought, they suffer from hunger. Lack of food is the immediate cause of their distress, though lack of water may be the cause of the lack of food. The importance of water—via food—may be seen from figures 1–4. The correspondence between annual rainfall and density of population is very close, the chief exceptions in the old world being that the valleys of the Indus, the Tigris-Euphrates and the Nile are heavily populated in spite of having a very low rainfall, while the Himalayas and parts of Central Africa are sparsely populated in spite of a good rainfall. The reason for the low population of the Himalayas is obvious; in Africa less so, though one of the factors may be the tsetse fly as this is the northern border of the area of sleeping sickness.

In North America the correspondence is also striking, though in this new country the population is relatively smaller in comparison with the rainfall. In South America the Amazon valley constitutes a rather striking exception, having a high rainfall and a very low population. Other resources, ores, oil, coal and even forests are useless for supporting human life unless they can be exchanged for food or can contribute to its production.

The amount of food and the number of mouths have concerned thoughtful people throughout the centuries, but efforts to define exactly the relationships between them date only from Malthus and are few in number. He, more than any other person, brought out the concepts that resources are limited, that all biological populations tend to grow without limit, and that man is no exception to this general rule. He furthermore pointed out that it is the ratio of food to mouths that is important and not the absolute values. There is little use in considering one without the other. Familiar as these concepts are, many people, even biologists, seem to forget them and to talk as do the politicians of shortages of food when they mean relative excesses of people. The amount of food in the world has been increasing for the last three or four centuries. How then can we speak of a shortage of food? It is the excessive increase of the population, the denominator of the fraction, that has caused dislocations of the ratio.

Consider for a moment the earth as a total organism: Will the cosmic physician, looking upon Mother Earth as a woman and mankind as an enlarged thyroid causing excessive metabolism, exclaim "There is a disproportion between this woman and her thyroid; we must enlarge the woman," or will he perhaps consider irradiation of the gland?

There are two reasons for discussing India in this connection. The first is that in India you may observe on a large scale what may be called 'natural agricultural man,' that is, man whose relation to his environment is the same as it was in the great river valley civilizations of the past, in Egypt, Babylonia, India and southern China, indeed, in all the great civilizations of recorded history. Furthermore, it is still the most prevalent relationship. The majority of the inhabitants of the world live under conditions approximating those in India and only a few under conditions approximating ours. We tend to forget that our type of life is recent, exceptional, and restricted to a small area which is probably shrinking as the population of the world outstrips its food production. The relationships observed in this country are not typical of the world at large, and as Bertrand Russell has recently pointed out, there is no guarantee that they will last (1).



Figs. 1–4

Let me illustrate what I mean by considering a young ecologist who is interested in studying the relationship between man and some other organism in the environment, let us say tigers. He may start his research by going to Ringling Brothers' Circus and sitting immediately outside the cage in which the trainer is putting the big cats through their paces. In this way he can get very close indeed to the tigers and watch and even photograph their every move as well as observing the man. He will, however, come away with a very incomplete and one-sided impression of the ecological relationship between men and

tigers. He will learn about the man-tiger relationship which is characteristic of the United States, but most men and most tigers do not live in the United States.

The same young ecologist going to India will find that there is not the capital there to provide a trainer and an expensive cage for each tiger and to provide him with nourishing food conveniently served. That makes a great difference. He will perhaps not even see a tiger. Nevertheless, he will, in the tiger districts, get a far more exact and vivid impression of the 'natural' man-tiger relationship than he could acquire here.

I have labored this point because I want to make very clear what I mean by 'natural ecological relationships.' I mean those that have existed throughout history, that apparently tend to become established whenever large groups of men have settled for long periods in a given area, and that are predominant in the world today. I also wished to point out how completely the investment of large amounts of capital could change the ecological picture, and finally I wanted to indicate the desirability of field experience. You may catch a *plasmodium vivax*, perhaps host and all, and bring it to this country for study, but until you go to its country for a study you will not understand the malaria problem.

The second reason for choosing India for discussion is allied to the first. It is that India demonstrates a fully developed and stable culture. The ecological relationships are not only natural; they are also mature. Here I will borrow a concept from forestry. When a piece of land in New England is cleared, farmed for a while and then abandoned, it gradually reverts to forest, but the first composition of the forest though a natural one is not the final one. Gray birch closely followed by white pine may take over initially, but after a hundred years or so these trees are pushed out by a variety of hardwoods. What is called a 'climax forest' gradually appears and is stable in composition until some great catastrophe, usually extrinsic, initiates a new cycle. India in my opinion represents a 'climax' ecology and culture. We, on the other hand, pretty clearly represent a first growth. India does not represent the only possible climax culture, but she does represent the one which seems to have been most commonly achieved in the past and most probable for the future. She shows the destination toward which the world as a whole appears to be moving.

She therefore merits the closest study by those groups of mankind who are still in an early stage of ecological development like the United States, where the ratio of time spent on getting food to total time is low. Many people have the impression that India is an 'under-developed' country while we are a highly developed one. Nothing could be more misleading than this 'quarter-truth.' We started at a higher stage of the arts and are exhausting our resources and increasing our population at a higher rate than India apparently achieved, but India went through the stage we are now in many centuries ago. From 1000 B.C. to 500 A.D. India occupied the position that Europe occupied from 1400 to 1900 and that we are now occupying, namely a vigorous, active, growing nation, exploiting their resources, expanding their culture and exporting their manufactured goods. The Roman emperors were much concerned by the rate at which gold flowed from Rome to India (in exchange for silks among other things) and was buried there as ours is today in Kentucky.

TABLE 1. GENERAL CHARACTERISTICS

	INDIA	U. S. A.
Food/Population, F/P	small	large
Time spent producing food/Total working time, T_F/T_T	90%+	25%
Birth rate/1000	46±	20±
Death rate/1000	35±	10±
Net increase/1000	11±	10±
Life expectancy at birth	28 yrs.	67 yrs.
Morbidity	high	low
Per capita income	small	large
Starvation	common	none
Social stability	great	small
Arts and sciences	inactive	active

Let us now turn from these philosophical concepts to actual observations, first rather general qualitative observations on the nature of India's mature civilization, and then to quantitative data on her population, area, food production and food imports. Both the qualitative and the quantitative observations will be compared with similar data from other countries in order to illustrate the determining influence of the food over population ratio upon human societies.

What are the characteristics of a natural and mature ecological relationship such as India has now developed? To my mind the most important ones are those shown in table 1, namely a high ratio of time spent on food production to total time, a high birth and death rate with low life

expectancy, almost everyone exceedingly poor, starvation common, stability great, and arts and sciences inactive.

Some of these characteristics need comment. Ninety per cent of Indians are rural and the great bulk work the land. Of the ones who do not actually work the land many are concerned with the transportation and handling of food. The figure commonly given for the number in industry is 4 per cent. In short, India is almost wholly an agricultural country and yet her most arduous efforts have failed to produce enough food to feed herself, and since 1943 there has been an increasing net import of foodstuffs. In the U.S.A. approximately one-fourth of the people easily produce food for the other three-fourths and export some besides. I will come back to this point later.

The next four figures obviously will vary from year to year and are here given in round numbers. The Indian figures are taken from Dr. Chandraseckhar's estimates published by UNESCO (2) and are about 30% higher than those found in the official statistics, as it is believed that nearly one-third of the births and deaths in India are not recorded.

The Indian life expectancy has varied considerably. From 1881 to 1911 it was about 25 years. It dropped to 20 years in 1921, rose to a high of 32 years in 1941, and now seems to be falling once more. These figures are uncertain.

To give the per capita income in figures would be more misleading than informative unless accompanied by a long discussion of comparative prices and cultural standards. Suffice it to say that ours is of the order of 10 times theirs at least.

Hunger is widespread always, and starvation in certain years is a significant cause of death. Within the last 10 years, namely in 1943, starvation caused about 20 per cent of all deaths in that year, and there is a definite possibility that during one of the next 10 years starvation will cause 50 per cent of all deaths.

The next item on the table that calls for comment is the social stability. This refers not to the froth on the top of society, which may be blown off at any moment, but to the state of life of the great masses. The British may come and go, Communism may come and go, but this scarcely touches the great masses. They work as hard as their starved bodies will allow and try unsuccessfully to resist the squeezes of nature and the fortunate few above them. It is that way now, has been like that for a long time and looks as if it would continue to be so. That is what I mean by social stability. It is a stable equilibrium but has no relation to individual security. Each life may be most insecure but the average never changes. In this country there is constant change and flow, living conditions change, types of work change, individuals rise and fall; in short, society is in a constant state of bubbling and flux.

The reasons for India's stability are fairly clear but are not often pointed out. In an agricultural country, when the combined efforts of the whole population are not quite sufficient to produce their minimum food requirements, no one can be spared from the immediate production of food to plan any changes for the future or to build capital goods. It is almost impossible to take a million men off the land to build dams, because if you do the food that they would have raised is not raised and therefore food for the builders of the dam must be taken from the farmers, who then die of starvation. The latest example of this process was in the Bengal famine of 1943, when the Indian government provided a large number of workmen for building roads, airfields, barracks and the like for the United States and British troops based there for the Burma campaign. These workmen produced no food but were fed on food bought locally, which raised the price until the farmers sold their own food stores, believing that they could buy food later cheaper. The price continued up and about 2–3 million farmers died of starvation. There were other factors in this famine, dislocations caused by the war, bad management, and also speculation in grain. In fact, in discussions of this famine the blame is usually placed entirely upon the speculators. However, it is hard to see that they could have had much influence upon it as they neither created nor destroyed grain. They did not export any; nor did they have any significant amount on hand at the end of the famine. No speculator wants to be caught with grain in hand in a falling market. If they did not alter the total calories consumed they did not 'cause the famine.' The employment of men on roads, barracks, airfields, etc. *did* decrease the calories produced, and the money paid to the workmen and by them to the farmers was not edible, and, due to the war time interference with transport, was not exchangeable for food. Most of the money wound up in the hands of the speculators, for starving people can never keep their money; they obviously exchange all they have for food. But there

just was not enough food, speculators or no speculators.

This brings us back to stable poverty. If no one can be spared from raising food to produce any capital goods, or large scale projects like dams, or to do anything but grub for food, it is clear that the situation tends to be static and stable. The arts and sciences tend to be inactive for the same reason, namely that so few can be spared to pursue them.

Let us for a moment consider the condition of a culture when only 25 per cent of the working time suffices to produce an abundance of food. Then three-fourths of the people are potentially free for any activity they may choose. They may choose unwisely and often do, but just the undirected more or less random activity of these free people in this country has produced the greatest amount of capital goods that the world has ever seen, and many of these capital goods, i.e. roads, tractors, farming machinery, processing machinery, scientific investigation and knowledge, have very greatly increased the productivity of our already productive land. The effects are cumulative and cause great changes.

India too was once rich, powerful and expansive. What happened? What great disaster or gross mismanagement has reduced her to permanent poverty and starvation? There has been no great disaster. There has been merely the predictable operation of a natural law, a relentless rain of small disasters, 46 per thousand persons per year. This, with the help of better transportation and public health measures, has finally depressed the ratio of resources to population to its present level. And to any of you who may be shocked by my referring to babies as disasters, I can only say that I believe that field work in Calcutta would convince most people that when the F/P ratio is small enough, babies are both a personal and a general disaster.

Figure 5 gives a smoothed curve of the probable growth of the population of India and Pakistan together. All the figures have a considerable element of uncertainty which increases as you go back, so that the early periods are really pure conjecture. But from the remains of cities, from history, from burial grounds, and monuments and temples of all kinds, it is evident that the population was considerable as early as 1000 B.C. or even earlier. Most guesses are that for 2 or 3 thousand years before 1850 the population was in the vicinity of 50–100 million. The cause of the great increase between 1850 and 1900 cannot be defined with any confidence. It coincided with the improvement of internal transport by road and railroad, and with the organized efforts of the British to ameliorate the local famines by taking grain from surplus areas to the stricken ones. The curve in 1911 and 1921 looked as if it might level off at around 300 million as suggested by the open circles in figure 5, but the appearance of leveling off was in part produced by the influenza epidemic. It is probable that the steep increase since 1921 is in large part due to better control of the great epidemic diseases, smallpox, cholera and plague.

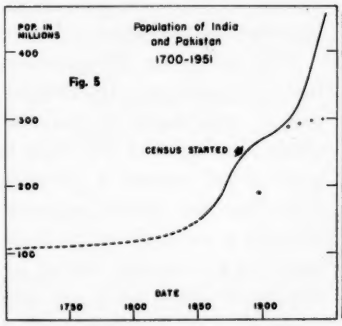


Fig. 5

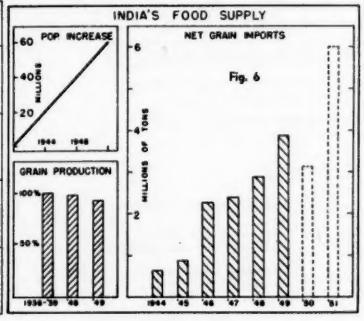


Fig. 6

The present situation in India may be summed up in the following (fig. 6): Since 1940 her population has been increasing at approximately 5 million per year. Her total food production appears to be declining, or at best remaining stationary. It seems to have reached its peak about 1940, and the years 1947–48 and 1948–49 showed 98 and 93 per cent of the pre-war average production, respectively. The discrepancy has been made up as shown in the net import balance in millions of tons. The figures for 1950 and 1951

are provisional. It goes up nearly a million tons per year. It takes about a million tons to feed 5 million people, and India now has at least 30 million more people than she can feed. For an agricultural country to have to import 10 per cent of her total caloric requirements is a disastrous situation. In India at present, as in many other places in the past, man has demonstrated that like other biological species he tends to multiply beyond the capacity of his range.

What can be done about it in India? The theoretical possibilities are many, but in my opinion most of them are not practicable. India has one-half our area and three times our population (I am including Pakistan in this figure) with a smaller proportion of good land. She cultivates

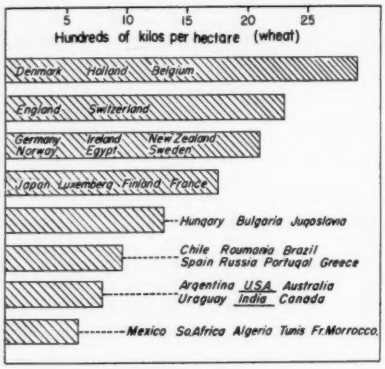


Fig. 7

over 50 per cent of her total and we under 20 per cent of our total, choosing naturally the best. There, there is a little over one acre of useable land per capita of the farm population, here about 12 acres per capita. One and a quarter acres—a piece of ground, often very poor ground, 75 yards by 75 yards—is too small an area to grow a year's supply of food, shelter, clothing and fuel for a man. Twelve acres of good land is plenty for four men. That is the situation in its simplest form.

It is true that it is possible to grow more on each acre than the Indians do. I suppose there is not a single acre in the world that will not grow more than it now does if enough capital, labor, and knowledge are put into the effort, but capital and knowledge are the accumulations from past surplus and India has had no surplus for many years. It was all used up long ago.

Figure 7 shows the productivity of various countries in deci-tons per hectare for the period just before the war when the iron curtain did not divide Europe. India and the U. S. A. are near the bottom of the list. We are now teaching India how to produce more. It looks like the blind leading the blind and perhaps it is so to some extent, but it is less so than it looks because the reasons for low productivity are opposite. In India, the low productivity is mostly due to poor land and too little water, and partly to inefficient methods of utilizing what land there is. Here, it is due to the fact that there is so much good land, and that extensive farming produces more grain per man hour than intensive farming, such as Denmark practices. Here man hours are expensive, in India they are a drug on the market. Currently the Indian Government is trying to put an artificially high value, namely 6 cents, on the industrial man hour, but the natural and historic level of an agricultural man hour is about 2 cents or less.

This brings up the question of the introduction of machinery. In discussions of the usefulness of machinery in India we in this country often lose sight of the fact that it will do no good at all unless it produces *more* food per acre than the present methods. Furthermore, if you get a machine which takes the place of 100 men, these men will starve to death as soon as displaced unless they are either *1*) given food while they do nothing, or *2*) put on other jobs and given salaries. This latter course requires capital which India has not got. That is why Mr. Hughes' current efforts on the Point Four program are so excellent, with their emphasis less on the introduction of machines and more on the improvement of yields by better practice with their current equipment.

But to my mind the question is, can you decrease the ratio of T_F to T_T, time spent on food to total time available for work. If you do not decrease this ratio you achieve very little, because India will not accumulate capital and will continue to be dependent on outside help. Furthermore,—and this is the most important point of all—if you do not stop immediately the increase of population you will get nowhere, because at her present rate of increase India's population will double in 50–60 years. India is stable as she is. It will take a huge push to get her to the point where she can again accumulate fat. It seems unlikely that such a push can be given unless the population is decreased, and countries that might give this push are few and shrinking in number. Indeed the size of push that India needs can come only from a large,

rich, and uninhabited continent placed at her disposal.

I will close by discussing some of the very few observations that I know of on the relationship between calories fed and production obtained in humans under natural (as opposed to laboratory) conditions. This relationship has been very well worked out in hens, horses, cows and other domestic animals but we have been slow in recognizing how completely it also may apply to man. If it takes 10,000 calories to feed a horse which is not working at all, it will take about 30,000 calories to feed the same horse at hard work, producing say 6 horse-power hours per day. It is obviously more economical to feed that one good horse 30,000 calories than to have 6 horses each doing 1 horse-power hour of work per day but eating, in the aggregate, 80,000 calories. India is following the latter course with her man power.

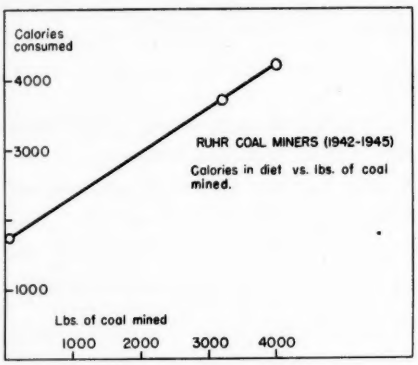


Fig. 8

This same principle can be illustrated by the data from the Kaiser Wilhelm Institute upon the Ruhr miners during the war (fig. 8). As this was not a planned experiment the data is far less complete than one would wish, but it gives some very suggestive information. It was found that when they received 4,200 calories per day, which was their normal average, they produced 2 tons of coal apiece. When during the shortages caused by the war they were cut to lower diets, their output fell off on an approximately straight line which indicated that they would reach zero production at about 1800 calories. Exhortations to the men did not affect the production for very long, though they did produce a little loss of weight.

Thus, feeding these men 1800 calories per day would produce no coal but every 100 calories over that would produce 167 pounds of coal up to a maximum of 4,000 pounds. The biological importance of this concept may be appreciated from the following consideration. Let us suppose that ordinarily 4,000 pounds of coal will purchase 4800 calories of food. Each miner uses 4200 calories and gets 4800 calories for his 2 tons, the 600 extra being exchanged for other goods. He gradually accumulates wealth. But let an interruption occur so that the food available is only 3,000 calories. His output will at once fall to 2000 pounds, whose purchasing power is only 2400 calories. This will produce only 1000 pounds and this will only buy 1200 calories, which is too little to live on. The situation has elements of instability in it.

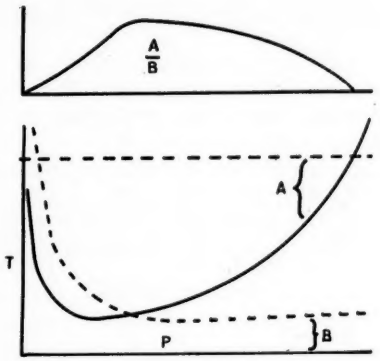


Fig. 9

The final figure (fig. 9) is entirely theoretical. It shows the kind of curves that I would like to see developed for various populations in various areas. Such graphic representation would be of great help in objective thinking on many human problems. Three curves are presented, the coordinates for the lower two being working time, T, and population, P. The solid line on the lower set of coordinates represents the fraction of the total available working time which must be spent on food at a given state of the arts. The horizontal dotted line represents 100% of the working time. At first, as the population increases production becomes easier. The forest can be kept back with less difficulty if many miles are cleared than if there is only one small clearing. Later, as the good level rich land is all taken up, the number of man hours that must be spent on food increases. The dotted curve gives the cost in man hours per unit of consumer goods, refrigerators, for example. This falls very rapidly at first because of the economies of mass production and

then stays level, and may even rise very slightly as the project becomes huge and costs of organization rise.

The top curve is the ratio of A to B, namely, the amount of time left over from the production of food divided by the cost of a unit of consumer goods plotted against population. This ratio really represents the standard of living. It shows the number of units of consumer goods other than food that the individual can have. The shape of this curve is interesting, particularly the tendency to concavity upwards in the initial portion and to convexity (that is accelerating decline) in the latter portion. This is an indication of instability. Once you get well onto this decline it is hard to stop from sliding down the rest of the way.

CONCLUSION

I have endeavored to show that the ratio of resources (and especially food resources) to population largely determines the whole nature of a civilization and that this ratio is now low and sinking not because of a shortage of food, as is often said, but because of a relative excess of population. I have suggested that when this ratio is low the situation tends to be a stable one of permanent poverty for most individuals terminated by early death. I have given examples of how general biological laws may apply very exactly to men. I now wish to close upon the same note on which I began, that these problems in human ecology are of the greatest magnitude and urgency. They call for immediate and objective analyses by trained biologists.

REFERENCES

1. Russell, Bertrand. *Brit. Med. J.* Dec. 10, 1949, p. 1307.
2. Chandraseckhar, R. *La Démographie de l'Inde et Pakistan*. UNESCO, 1949.
3. Davis, Kingsley. *The Population of India and Pakistan*. Princeton: Princeton Univ. Press, 1951.
4. *Statistisk Årsbok för Sverige 1948*. Stockholm: Kungl. Boktryckeriet, 1949.
5. Brody, Samuel. *Bioenergetics and Growth*. New York: Reinhold Publishing Co., 1945.
6. *Records and Statistics*. (A quarterly published by the Eastern Economist, New Delhi, India.)

AGRICULTURAL AND INDUSTRIAL POSSIBILITIES OF INCREASING THE FOOD SUPPLY

L. A. MAYNARD

From the School of Nutrition, Cornell University, Ithaca, New York

SPECULATIONS as to how much more food could be provided if all the potentially available land were utilized, if our forest resources were transformed into food with the aid of chemistry and microbiology and if our global waters were 'cultivated' as intensely as our land areas, fire the imagination. A more realistic approach is needed to the question: "How many people can we feed?" Economic, social and political factors, as well as basic resources and technical 'know-how,' govern the extent to which food production can be stepped up. These factors also determine how successful a program can be in making more food available to those who need it most. How many people we might theoretically feed and how many of them would actually eat their share are different questions.

There is also the question as to the kind of diet we are thinking of. In the United States we could produce 3000 calories a day for over three times as many people as we are now feeding, by using our grain crops for direct consumption and limiting our livestock to the numbers which could be produced on grasslands alone, with no grain feeding. Of course, such a radical change is out of the question, for many reasons. The situation for a vast number of the world's population lies at the other extreme. Their diet consists so largely of cereals or other carbohydrate goods, under the primary urge to get enough to eat, that its nutritional quality is seriously deficient. It is not enough to think merely in terms of calories or of tons and bushels of food, in estimating how many people we can feed. The food should be of a kind which will meet at least minimum nutritional standards in terms of health and performance.

I shall review various possibilities for increasing the food supply, with especial attention to factors which will govern the extent to which these possibilities can be achieved. I am aware of my rashness in attempting this assignment, particularly in view of the many variables to which any quantitative estimates must be subject. For example, much depends on the outcome of current and future research. Scientists realize that the extent of their contributions cannot be foretold.

At the outset, it must be recognized that we are fighting a losing battle in feeding the world at the present time. The latest report (1950–51) of the Director-General of FAO states that, though world production is now 9 per cent greater than the average for the period 1934–38, the number of people have increased by 12 per cent. Per capita consumption is down and the situation has worsened most in areas which were previously critical. One might blame political unrest for the losing battle, but the future holds no immediate hope for a more favorable situation. We cannot let the past lack of success unduly discourage us, but it should make us more conservative in predicting what can be accomplished in the future. It should also suggest the lines along which a more intensive effort for increased production might be most effective.

AGRICULTURAL POSSIBILITIES

The agricultural possibilities for increasing food production fall into three classes: *1*) a larger and more efficient production on land already in use, by taking fuller advantage of practices of proven superiority, *2*) the utilization of land now wasted, and *3*) the discovery of better practices through scientific research.

It is agreed that a large opportunity for increased food production lies in the application of improved practices to the land already in crops. Under the incentive of market demand and aided by favorable weather, the U. S. A. food production increased by 40 per cent from the 1935–39 period to 1945, mostly as a result of increased yields per acre. A food supply of higher nutritional quality also resulted because of the relatively larger increase in animal products, fruits and vegetables. The use of superior strains of plants and animals, larger fertilizer applications, better farm tools, new methods of disease and pest control; and better breeding, feeding and management practices with animals, all play a

role in this remarkable increase in food production.

Agricultural experts recognize that these improved practices have by no means been put into full operation across the country. Groups of scientists in our experiment stations, Department of Agriculture and other government agencies have been engaged in making a detailed and careful estimate as to how much our present production might be stepped up by taking full advantage of present 'know-how.' From an examination of some of the preliminary results of this overall study, Salter (1), formerly Chief of the Bureau of Plant Industry, Soils and Engineering and now Chief of the Soil Conservation Service, concludes that present agricultural production could be stepped up by 60–75 per cent if the best combinations of known practices were put into use on all farms. He recognizes that this is an ideal which would not be fully achieved. It would appear that one-half of this increase would be a much more likely possibility of achievement over the next quarter century, and this would require favorable economic conditions. Nevertheless, Salter points to the fact that our best farmers are now making their soils produce nearly twice as much as the average farmer is accomplishing. He mentions that, although the national average corn yield is less than 40 bushels per acre, more and more farmers are getting 100 bushels each year.

Table 1. Estimates of yields/acre attainable by 1955 (Indiana)

Crop	Yield/acre (bushels) Av. for base period, 1946–50	1955 Attainable	1955 Maximum
Corn	50.9	63.5	91.6
Wheat	21.9	27.1	34.6
Oats	37.7	52.2	61.3
Soybeans	21.1	24.2	33.2

The data being obtained in this nation-wide study are illustrated by those reported by the Indiana Committee on Agricultural Production Capacity (2), for a State which has already become one of the best yielding areas of the country. The estimates made for the 4 major crops are reproduced in table 1. The figures listed as 'attainable' are the yields which could be reached by 1955 by practices which would be most profitable and best suited to the farmers' resources under favorable economic conditions. The 'maximum' concept assumes the full adoption of all known improved practices under skilled management. It is significant to note, for example, that, while the corn yield is already much higher than the national average, a nearly 25 per cent increase is estimated to be attainable by 1955, with a maximum possible increase of approximately 80 per cent.

The United States is only a small part of the world. What about the possibilities elsewhere? Salter (3) has concluded that, while significant increases from the highly developed European countries are doubtful, there are large possibilities in other areas. Based upon estimates of increases attainable in the U. S. A. from similar soil groups, and taking account of present cropping practices in the areas in question, he has computed possible increases in yields, by 1960 compared with pre-war, for the major food crops in various countries. As an example, his figures for rice and wheat for China and India are reproduced in table 2. Increases from 30–50 per cent are here represented. Obviously these estimates are subject to challenge. Agricultural scientists agree that there are many areas of the world where yields per acre can be stepped up by the adoption of better practices, but that the bases for quantitative estimates are very meagre.

Table 2. Estimated attainable increases in yields/acre (Salter)

		Yield 1935–39 *bu*	Yield 1960 *bu*
Rice,	China	52.5	70
	India	26.2	40
Wheat,	China	14.9	18
	India	10.7	20

I've been talking about possibilities. Their realization depends upon education, capital, and an expanded industry to supply more fertilizer, chemicals for pest control, and farm tools. Favorable economic conditions would also be required. These various essentials must be available in the producing areas throughout the world and particularly where the people to be fed are concentrated, involving situations vastly different from the U. S. A. The present situation in the countries of greatest concern does not justify much optimism that these essential conditions can be met.

Under the most favorable conditions, expanding the use of the present 'know-how' will take time. In the meanwhile the population will con-

tinue to increase. Even here in the United States, many visualize a real problem in expanding our production fast enough to take care of the Census Bureau's estimated 190 million people in 1975, in terms of our present dietary pattern, unless we are much less wasteful of our food supply. This fact was brought out clearly by data presented by Shaw (4), Agricultural Research Administrator, at a recent Congressional hearing. He also stressed the necessity for increased research and soil conservation activities, if future food needs are to be met. In fact, if these activities should slacken, a declining food supply would result.

An increased production per man coupled with economic developments which would enable the released manpower to be utilized effectively in other enterprises would raise the standard of living and mean a better food supply for many. In 1900 each farm worker in the United States produced enough food and fiber for himself and about 5 other persons. Today each worker feeds 14 others besides himself. Farmers are now producing about 90 per cent more corn and 80 per cent more wheat per hour of labor than they did 15 years ago. Of course, this has involved a mechanization that we cannot visualize for many of the critical areas of the world. From my experience in China in the thirties, however, I am convinced that better hand tools and better agricultural practices in general can significantly raise the output per man on the small farms which characterize the areas of the densest population.

UTILIZATION OF LAND NOW WASTED

What are the possibilities of bringing additional land into production? According to data published by FAO (5), less than 10 per cent of the total world land area is under cultivation, though twice this percentage, as permanent pasture and meadows, is utilized to produce animals and thus human food. Much of this pasture or meadow land could be made much more productive by appropriate improvement procedures, such as fertilization, weed and brush control and other measures. There is also much land throughout the world, now entirely unused, which could be brought into production, under appropriate conditions. The FAO table giving the breakdown of the total world land area, lists approximately 950 million acres, as "unused but potentially productive." Salter's study (3), involving consideration of the classes of the soils in question, arrived at a figure which is one-third larger. Probably his estimate included some land classed by FAO as forest. On the basis of these two sources of data it appears that the present food producing land, exclusive of meadows and pastures, might be increased by one-third. Salter recognizes that how much of this area could actually be put into use would be determined by production costs in relation to returns. Clearly there would also be important social considerations involved.

Much of the unused land is naturally less fertile than that now farmed and heavy fertilization would be needed accordingly. Much would also require drainage, or irrigation or erosion control. Large initial capital expenditures would be needed. Many feel that the actual utilization of much of this waste land for food production will continue to be economically impossible. We do have examples, however, of such developments. Mexico brought a million acres under irrigation and improved the irrigation on another million acres between 1926 and 1946 by spending a billion pesos, according to DeGraff (6).

INCREASING AGRICULTURAL PRODUCTION THROUGH RESEARCH

The large increases in food production since pre-war have resulted primarily from the application of scientific discoveries made over the preceding years. These developments are of special significance in indicating the large possibilities which science holds for increasing our future food supply. More than four-fifths of our corn acreage is now in hybrids, which outyield the open pollinated varieties by 25 per cent. Corn is only one food crop. Certainly many more important contributions to our food supply, from harnessing nature's forces represented by the hybrid vigor, lie ahead.

Advances made in combating crop diseases are illustrated by the data indicating that the use of DDT has been responsible for increasing potato yields in the U. S. A. by 25 per cent. Even more important, particularly on a long-term basis, are the achievements made in breeding disease-resistant varieties of various food crops. The toll from thousands of diseases of economic plants is still measured in hundreds of millions of dollars annually in the United States alone. Much has been accomplished in the field of animal health, yet the estimated annual losses

from farm animal diseases in this country are valued at half a billion dollars. No one can deny that there are large possibilities of increasing our actual food supply and decreasing its cost through scientific research to find ways of saving the large losses now resulting from plant and animal diseases.

The recent discoveries in the field of nutrition have developed practices which have made the production of human food by farm animals much more efficient. Ten years ago it took about four pounds of feed per pound of gain to produce a broiler. Today less than three pounds suffices. In the past two years a further advance in this field has come from the feeding of small amounts of certain antibiotics. By their use, rates of gain in pigs have been increased by 10–20 per cent and in chickens by 5–10 per cent, with a decrease in the total feed required. One cannot visualize, for most of the world, any large consumption of animal products. But a certain amount is clearly desirable in the interest of dietary quality. Any discoveries which make the animal more economical in the use of its feed should help in this. Synthetic hormones are being employed to step up animal production and also to regulate plant processes, with resulting larger yields. The development of hormones for combating certain plant diseases is also an active field of research. The recent accomplishments in the chemical control of weeds and brush are also of large significance.

I have merely given a few examples of what science has and is currently accomplishing in increasing food production, to indicate its future promise. It should be stressed, however, that the largest possibilities lie in basic research in the sciences underlying food production, with no more immediate goals than the accumulation of knowledge. Solving the many riddles which now exist regarding the chemical and biological processes which take place in plants and animals may be expected to lead to greater efficiency in food production and a larger supply accordingly.

POSSIBILITIES OF INCREASING THE SUPPLY OF FISH

Seventy per cent of the earth's surface is not land but water. What are the possibilities of increasing the food supply from this source? This is a question which has received special attention on a world-wide basis by FAO. This organization estimates that the current production of fish is approximately 25 million metric tons annually and that this production could be doubled without risk to future fish resources. What does this mean in terms of feeding the world's population? Fish are of special dietary importance because of the high quality of the protein furnished. FAO estimates that two-thirds of the present total production is available for human consumption and that 40 per cent of the latter is actually edible. Applying these figures to the known data as to the yields of the different species and as to their composition to an estimated world population of 2 billion, 400 million, it can be calculated that the present fish supply is furnishing approximately 10 calories and 1.3 gm. protein per capita per day. Thus, even doubling the present supply would not make any large contribution on a world basis. In many areas adjacent to the sources of supply, however fish now form an important part of the diet. In such areas, as well as in similar ones where the fishing industry has been little developed, the possibilities visualized by FAO have a real significance.

Apart from doubling the fish supply as now marketed, there are other possibilities. More of the species now considered inedible might be utilized and the total catch might be marketed with less wastage. For example, according to the United Nations Interim Commission on Food and Agriculture, over 100 million pounds of fish taken annually by trawls off the New England coast and in the Gulf of Mexico are discarded as being unmarketable. Many million pounds of shark carcasses, suitable for the production of salt fish, are returned to the sea after removal of the vitamin-rich livers. New sources of supply not visualized in the present FAO estimate may be found. Certainly we can look forward to increased food supplies from our global waters, but the suggestion that in the future sea farming will be comparable to land farming, with large fenced-off water areas, doesn't seem very realistic at present. Contrary to popular ideas, the sea is not full of fish.

Fisheries development requires capital, specialized equipment and trained technical personnel. It is bound to be slow. Basic research and pilot projects are particularly needed at many points. Per ton of food, we are spending much less for research on aquatic sources than on land production. One could expect that a stepped-up research program would develop procedures for increasing the supply of fish in ways not now visualized.

The pond culture of fresh-water fish, which

has been practiced to a certain extent in Europe for many years, is now being actively promoted for much wider use. In specific situations this development may result in an important contribution to the food supply of individual families and perhaps of communities. But, for various reasons, any significant contributions to the world food supply cannot now be visualized.

INDUSTRIAL POSSIBILITIES

Industry has obviously played a large role in the agricultural developments which are stepping up food production. It has produced more and better fertilizers, designed and built better farm tools and machinery, and developed the large-scale production of the new chemicals for disease, pest and weed control.

The remarkable progress made in the cheap, commercial production of certain vitamins and the beginnings made in the field of amino acids deserve note in considering future possibilities. The food supply of many peoples of the world is deficient in some of these nutrients. In some of these areas it may appear essential to continue the same general food production pattern in the interest of obtaining the largest possible yield of calories. Here a practical way may be found of supplementing the diet with certain deficient vitamins and amino acids. The success of the rice enrichment program now in operation in the Philippines and elsewhere in the Far East is a case in point.

Larger industrial possibilities lie in the conversion of inedible plant materials into useful foods. It is in this connection that our vast forest resources are mentioned. There have been many wild speculations in this area but the actual possibilities deserve analysis. In World War I in Germany and Scandinavia, wood and various other cellulosic materials were hydrolyzed to lower carbohydrates which found use as cattle feed. In World War II the process was extended to provide 'wood sugar' which was used directly as human food. The feasibility of the process was demonstrated as an emergency measure where the raw materials were near at hand and the needed industrial plants and chemicals could be provided. One cannot visualize much promise in such a development in the areas of the world where the food problem is chronic and becoming more acute as population increases.

A wartime development which has received continuing attention was the production of food yeast from wood sugar, as well as from various carbohydrate wastes. On a dry matter basis, yeast contains approximately 50 per cent of protein which is of high biological value. This protein can be produced from inorganic nitrogen compounds which man cannot utilize, an inexhaustible supply of which can be obtained from the air. Yeast is also a rich source of certain vitamins which are deficient in many commonly used foods. Wood sugar can provide the energy needed for its growth. Thus, the potential importance of wood yeast for increasing the food supply lies in its utilization of vast resources otherwise unavailable for man's consumption, particularly as a supplement in areas where the crops providing the largest yields are deficient in protein and vitamins.

At the close of the war six plants with a total annual production capacity of 25,000 tons were in operation in Germany and other plants were under construction. The yeast was used in feeding both people and animals. Attempts since the war to develop food yeast production on a large scale, notably in the British Commonwealth, have not been successful. Costs of production, and problems of getting the yeast produced and consumed in the areas where it should be of most benefit, have proven deterrent factors. The prospects for markets in the immediate future have not been such as to supply the capital and scientific interest needed for attempts to solve the problems met in the development thus far. These considerations should not rule out food yeast as a future possibility.

There are obviously other microbiological possibilities of increasing the food supply. In this area, a recent scientific discovery with *Chlorella* is attracting wide interest. The growth rate of this green, fresh-water, unicellular organism can be multiplied ten-fold by enriching the concentration of carbon dioxide in the culture flask, which in turn speeds up photosynthesis. Of greater scientific interest is the finding that by varying the cultural conditions large changes in the percentages of fat and protein in the organism can be brought about. The design of equipment for mass production of *Chlorella* is under study, but this is only one of the problems to be solved in making such a potential food source available in a useful and acceptable form for human consumption. Actually, the basic importance of the scientific findings may have much greater significance in the leads they provide for harnessing more of the sun's energy for food production in general.

CONCLUSION

Any review of the possibilities for increasing food production, such as has been made in this paper, is subject to challenge as being speculative. I wish I could summarize them in terms of their realities, but this cannot be done. Certainly one should take a conservative position as to what is likely to be accomplished in the near future, particularly in the light of the conditions which we can foresee. In the meantime, the world's population continues to increase. The possibilities I have discussed provide no justification for any complacency regarding this problem. On the other hand, the past and current developments in food production, and particularly in the scientific research underlying them, should cause us to keep an 'open mind' as to how much more might be accomplished in the future along the lines here reviewed. Certainly scientists should be willing to do this. The years to come will undoubtedly produce many more papers dealing with this general topic.

REFERENCES

1. SALTER, ROBERT M. *Address at National Farm Institute*. Des Moines, Iowa, Feb. 15, 1952.
2. *Report No. 2 of Indiana's Committee on Agricultural Production Capacity*. Lafayette: Purdue Agr. Expt. Sta. and Extension Service, 1951.
3. SALTER, ROBERT M. *Science* 105: 533–538, 1947.
4. SHAW, BYRON T. *Testimony before the House Subcommittee on Agricultural Appropriations*, Feb. 5, 1952.
5. *Yearbook of Food and Agricultural Statistics, 1950*. Washington: FAO, 1951.
6. DEGRAFF, HERRELL. *J. Home Econ.* 43: 411, 1951.

FACTS, FABLES, AND FALLACIES ON FEEDING THE WORLD POPULATION

SAMUEL BRODY

From the Department of Dairy Husbandry, University of Missouri, Columbia

I think we, in biology, must acquaint our fellow citizens with this serious world problem and aid them in working out a wise solution.

ANTON J. CARLSON (1)

THERE is no food shortage in the USA. Our preoccupation with the world food shortage is due to explosive political situations and menacing revolutionary fury in other parts of the world, particularly Asia, because "two-thirds of the world's people, even before the war, were undernourished all the time" (John Boyd Orr, first Secretary-General of the FAO); and "marginal millions do not get enough to eat, which means a race with death" (Norris E. Dodd, present Director-General of the FAO).

The remedy for food shortage is to increase food production. This is being done. The population in the densely populated areas is increasing, however, more rapidly than the food production, as shown by the 1951 FAO report (2):

> Most of those who were hungry in the five prewar years are now hungrier. . . . Production in most of the undernourished areas is failing to keep pace with population growth. . . . This decrease means more undernourished. . . . less resistance to disease . . . more funeral processions of those who die young. . . . Hunger is steadily haunting our civilization. . . . There will inevitably be also a rising tide of unrest and revolution and war.

Furthermore, most of the world's accessible arable land is being cultivated for food production and there are limitations on the ability of the earth to feed indefinitely a free-reproducing population. Therefore, an attempt should be made to reduce the pace of the population increase.

The problem of population-growth decrease is, however, unlike food production increase, bedevilled by numerous traditional and emotional taboos and confounded by many confusing fables and fallacies. Public speakers, especially natural scientists, avoid getting involved in such complications, preferring to keep their problems simple like their laboratory researches. Unfortunately, soft-pedalling the population factor in the food/population ratio does not advance the overall solution of the problem. Hence, the following hazardous attempt to treat this situation in its broader aspects—including the population factor—even if it does involve mixing natural science with social science. The necessity for such broad treatment is due to the fact that the behavior of man is profoundly affected by ideas of his own making. His religion and his social science influence his pattern of action even as much as do purely biological factors (3).

POPULATION GROWTH

It is a basic biological principle that "there is no bound to the prolific nature in plants and animals but what is made by their crowding and interfering with each others means of subsistence" (T. R. Malthus, 1798). How does this principle apply to human populations?

1. Historical perspectives. Most Occidentals have the impression that the food shortage in Asia is due to the high percentage increase of its population. This is not true. The *percentage* growth of the population in Asia, with respect to the people there, is less than in the USA (fig. 1). The population growth in Asia, however, is extremely great *with respect to the available fertile land.*

This brings up an historical event perhaps overlooked. It is that 15th century Europe was in the same cramped condition that Asia is in now. Then something happened to relieve it: Columbus discovered America, a territory 6 times larger than Europe, with a great abundance of fertile soil and rich minerals. Australia, New Zealand, South Africa, and the many newly discovered islands belong in this category of free fertile land. These territorial discoveries rejuvenated old Europe, gave her the needed elbow room.

The new territories not only had free fertile

land and other riches, but an invigorating climate—air free from kaisers, czars, kings, secret police, inquisitions, a situation unprecedent in human history. Here a novel doctrine was born[1]: "All men are endowed by their Creator with certain inalienable rights—life, liberty and the pursuit of happiness," the foundation of the 'American way of life,' which includes the Canadian, Australian, New Zealand, and most Western European ways of life. This way of life is now in danger because of the decline in food *per person.*

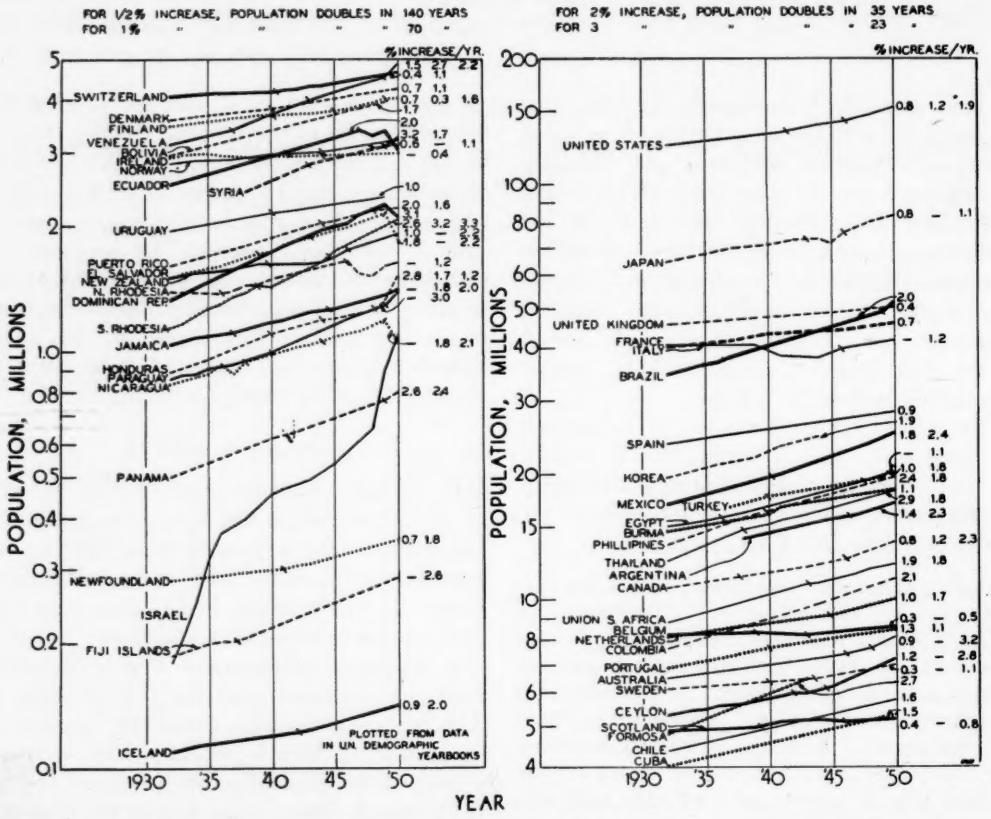


Fig. 1. POPULATION plotted on arithlog grids against calendar years, 1932–1950 (from data in *UN Demographic Year Books*). Crossbars represent approximate positions of change in slopes of curves. Slopes of curves × 100 represent annual percentage increases in population; i.e., USA curve 1932–1950 has 3 segments of slopes, 0.008, 0.012, 0.019, representing annual percentage increases of 0.8, 1.2 and 1.9 per year. For a 1% slope population is doubled in approximately 70 years; for 2%, in 35 years; for 3%, in 23 years; for ½%, in 140 years. (Time doubling population = $0.693/k$; thus for annual 1% increase, population is doubled in 0.693/0.01 = 70 years.)

Cause and effect cannot be disentangled, but the rise in modern science and technology, beginning with the Industrial Revolution, paralleled the rise in territorial discoveries and particularly their development. A major result of the scientific progress and technology was development, first slowly then more rapidly, of public health control methods with associated increases in life expectancy. These began with governmental control of communicable diseases, and culminating in the modern large-scale manufacture of chemotherapeutics, antibiotics, and pesticides which virtually eliminate death from infection, and increased the average life expectancy from 30 years in the fifteenth century to over 70 years in contemporary USA. The rise in life expectancy was particularly rapid in the past 50 years (fig. 2).

This life-expectancy increase was accomplished mostly by increasing the survival rate of children who, in turn, lived to reproduce, so that the population tended to increase *autocatalytically* (exponentially, in compound-interest fashion)

[1] Actually this doctrine was not 'born' here, but an opportunity was given for its development. It goes back to the Bible, and to John Locke (1632–1704) who spoke of "Life, liberty, and property".

with time, with increasing life expectancy. In a lesser extent these developments were made available to the rest of the world. The death rate was steadily declining while the birth rate was remaining roughly constant in most of the world, with a resultant rise in the ratio of births to deaths, that is in population (table 1, fig. 1).

In brief, the increasing world food shortage *per person*, reported by the FAO, is associated with an almost pathological upsurge in world population due to declining mortality and constancy of birth rate. I say 'pathological' because it is historically fantastic and, if not checked prophylactically, may affect the world population as cancer affects the individual.

To illustrate, let us assume that Adam and Eve set up housekeeping only 2100 years ago (even though it is believed that man has been on the earth at least a million years), and that the resulting population grew at the modest rate of 1 per cent per year (the present world population growth is 1.2 per cent per year as shown in table 1; see also fig. 1). The human race would then have reached in 2100 years a population, P, of 2.6 billions

$$P = 2e^{0.01 \times 2100} = 2.6 \text{ billion} \qquad (1a)$$

which is greater than the present world population of 2.4 billion.

To carry this computation to an absurdity, let us assume that Adam and Eve set up housekeeping 5300 years ago, and that the population increased at 1 per cent per year. In 5300 years, the population, P, would number

$$P = 2e^{0.01 \times 5300} = 1.0 \times 10^{23} \qquad (1b)$$

individuals, weighing (at 100 lbs. per person) 1.0×10^{25} lbs., equal to the weight of the entire earth. The earth's surface would then be covered miles thick with human beings. This is good arithmetical proof that the present rate of population growth is something really new under the

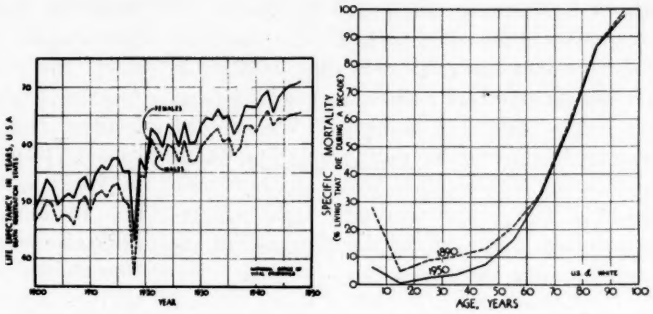


Fig. 2. LIFE EXPECTANCY at birth in USA during past 50 years increased by almost 50% (about 23 years) with corresponding increase in population growth for a given birth rate. (Right hand chart based on data by R. L. MCNAMARA, University of Missouri.)

TABLE 1

REGION	ESTIMATED POPULATION[1]				POPULATION INCREASE				PERCENTAGE DISTRIBUTION,[2] 1948–50		
	1936	1948	1949	1950	1949	1950	1949	1950	Population	Production	Ratio Prod. to Popul.
	millions				*millions/yr*		%	%			
Far East	1,072	1,175	1,187	1,198	12	11	1.0	1.0	54.5	32.0	0.5:1
Europe	372	390	394	397	4	3	1.0	0.8	18.0	23.5	1:1
Near East	108	124	125	127	1	2	0.8	1.6	5.5	4.4	0.8:1
Africa (ex. Near East)	127	147	150	153	3	3	2.0	2.0	7.0	4.7	0.7:1
Latin America	125	156	159	162	3	3	1.9	1.9	7.0	10.1	1:1
U. S. & Canada	140	160	163	166	3	3	1.9	1.8	7.5	22.6	3:1
Oceania	10.7	12.2	12.4	12.6	0.2	0.2	1.6	1.6	0.5	2.8	5:1
Total	1,955	2,164	2,190	2,216	26	26	1.2	1.2	100	100	

[1] Ref. 2, table 13. Figures refer to end of each year.
[2] Ref. 2, table 2. Percentage based on totals, excluding USSR.

sun; and if allowed to continue without planned controls, must be checked by the harsh natural methods that operate to maintain constant subhuman populations: starvation, disease, violence, and premature death. Which method shall we, humans, choose?

Anthropologists believe that preliterate and medieval populations were kept roughly constant by 'natural causes,' such as famines, epidemics, and wars. Carr-Saunders (4) reported wide-spread infanticide, abortion, and other crude methods for limiting offspring: "No matter in what quarter of the world we look wherever there are native races, one or more of these customs is practiced except where native customs have been destroyed by European influence." Wide-spread taboos against conception during lactation, which may continue for over three years, are important checks (5). The present unusually rapid population increase is due to the doubling of life expectancy during the past century.

2. Ideals and politics. In 1943, towards the end of the second World War, President Roosevelt gave the world his inspiring peace aim of world-wide "Freedom from Want," followed by the Hot Springs Conference, and the birth of the FAO of the UN.[2] These activities led to wide-spread awareness of the seriousness of the world food problem. I recall the editorial by our own Henry A. Mattill in *Nutrition Reviews* which declared that "the world cannot long exist in peace if half its people are well fed and the rest ill fed." The symposia sponsored by the American Association for Advancement of Science on "Freedom from Want" (6), by the American Philosophical Society on the "Possibilities for Increasing the World Food Supply" (7), and this symposium on "World Food Problems" were undoubtedly partly inspired by FAO stimuli and by the world-wide freedom from want ideal. This was followed by the Point Four program in President Truman's 1949 inaugural address "for making the benefits of our scientific advances and industrial progress available for the improvement and growth of underdeveloped areas," and the 1950 British Commonwealth's Colombo Plan to help India, Pakistan, Ceylon, Malay, and Borneo (8).

The latest FAO reports (2) show that their efforts and the activities of the Point Four program made great educational and statistical contributions. The food *per person*, however, continued to decline.

An interesting feature of these aids is that the world's hungry peoples do not appreciate them. Communists characterized them as propaganda—an old imperialism in new dress. They believe, following Hobson (9), that Western powers were traditionally interested in undeveloped countries solely for their own investment opportunities.

Professor Mukerjee (10) suggested settling Indians on unused submarginal lands in other countries. Indians are used to cultivating such land and to extracting a submarginal living. In brief, one could fill up the empty spaces in America, Australasia, and Africa with hungry Indians. Had not Europeans, particularly Englishmen, filled up empty spaces displacing natives in America, Australia, and South Africa? Could not Indians do likewise? This plan refused, Mukerjee would consider the declared freedom from want aim as political hyprocrisy, to counteract communistic propaganda. The Mukerjee solution, a solely verbal one, is cited only for its psychological interest, perhaps an indication of the shape of things to come, indicated by the following quotations:

"Why should we regard it of unquestionable benefit to have 400 million Indians in 1945 where there were 250 million in 1870? As they increase in numbers and as they learn to manufacture steel and the implements of war, will they not want more *Lebensraum*?" (11). "A heavily industrialized India backed up by such population pressure would be a danger to the entire world" (12).

It is not surprising for a shrewd and compassionate, as well as a publicity-conscious, world leader to suggest a reassuring world-wide freedom from want peace aim at a time when a hungry world is suffering from a murderous war. Wishful dreaming about food and peace is not unnatural for hungry people at war.

It is surprising, however, that the professional hunger fighters should believe that freedom from want can be achieved by simply "doubling the total food production." The real problem is, of course, doubling the food production *per person* in a freely-reproducing population. Judging by the FAO reports, and *equation 1* this cannot be achieved on a world-wide basis without the use of what Malthus called "powerful preventive checks" against population growth.

3. Nature of population growth. The FAO reports and *equation 1* above show that it is not

[2] Food and Agriculture Organization of the United Nations, including representatives of about 70 nations.

possible to achieve freedom from want on a world-wide scale by increasing food production without resorting to preventive checks. The population catches up too quickly with the increased food supply. This problem was anticipated by Benjamin Franklin (13) about 1750, developed in detail by Thomas Malthus (14) about 1800, accepted by Charles Darwin (15) about 1950, and recently confirmed (experimentally and mathematically) by the Raymond Pearl School (16).

The time curve of population growth is typically S-shaped (fig. 3). In the absence of environmental limitations the curve has a rising slope; the population tends to grow at a constant percentage rate, in proportion to the number of reproducing individuals. Reduction of space or food below a certain limit depresses the rising

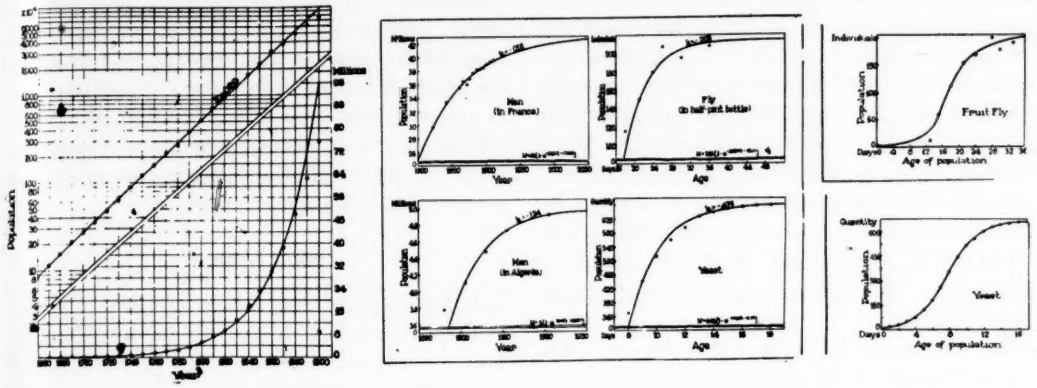


Fig. 3. TIME CURVES of growth of populations: yeast cells, fruit flies, man. *Right hand curves* illustrate sigmoid nature of population growth. *Left hand curve* illustrates initial exponential phase for growth of man in USA (*upper left* time curve on arithlog grid and *lower right* on arithmetic grid). Numerical value of slope, $k = 0.029$, or 2.9% per year; it means population was doubled in about $0.693/0.029 = 24$ years. *Middle curves* represent growth of two human populations following bend in the curve and effect of cultural evels of the two groups on their rates of approach to maximal population density. A new discovery may raise the ceiling, or begin new cycle from a higher base.

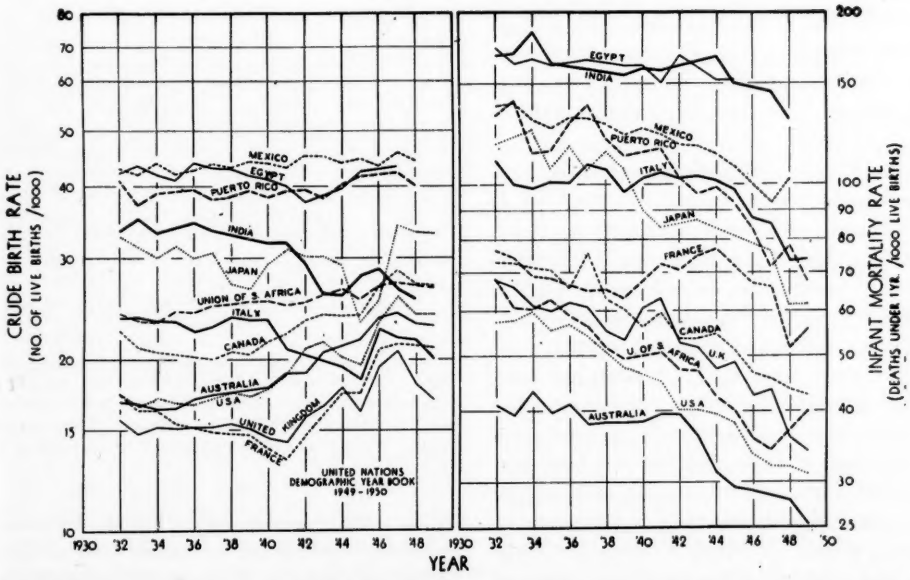


Fig. 4. HIGH INFANT MORTALITY is generally associated with a high birth rate. The birth rate in Egypt and Mexico is about 2- to 3-fold that in Australia and USA, but death rates are still higher so that these population increases are no greater than in Australia or USA (cf. fig. 1). The high death rate is evidently the price for the high birth rate.

course of the curve to a declining one, finally bringing the population growth to a standstill, to a dynamic equilibrium with the environment (17).

The curves of *Drosophila* in a pint bottle and of man in an isolated area *tend* to have similar sigmoid shapes (fig. 3) although the immediate limiting factors may be different. The shape and upper limit of the population curve of man depend not only on available food and space but also on his cultural level, traditions, and living standard, including deliberate family planning.

4. Population Data. The world population increased near 3-fold in the past 200 years, from 875 to about 2400 million. It is now increasing at the rate of 26 million or 1.2 per cent per year (table 1).

The population in India increased from 250 million in 1870 to 400 million in 1945 (11); it was increased by 83 million during the 1921–41 period (18). The population in Java was tripled between 1860 and 1930 (18). The population in Japan increased from 30 million in 1850, to 45 in 1900, to 56 in 1920, to 73 million in 1940 (18), and 83 million in 1950 (19).

The current increase in world population, 1.2 per cent per year, is not very different in Asia than in the USA (fig. 1); prior to 1900 the increase in the USA was near 3 per cent per year (fig. 3), higher than in any Asiatic country (fig. 1).

While the current population increase is, on a percentage basis, nearly the same in the USA as in Asia, the birth and death rates are very much higher in Asia than in the USA (fig. 4). Births in Asia are not controlled, except in the highest social category. In the USA, however, especially in the middle and upper-economic classes, conception is largely controlled. Children are born for desired parental pleasures—not unlike other things they desire and acquire—if circumstances permit. Because of their realistic attitudes and freedom from traditional inhibitions, and because of relative food abundance, there is no danger of 'Malthusian misery' in North America and similarly minded countries. Nor is there danger of serious population decline because most people are sufficiently fond of children to maintain the race even if they can easily avoid child bearing. The eager demand for babies by childless couples is good evidence that there is no danger of deliberate 'race suicide,' (cf. KAISER AND WHELPTON, *Studies of Psychological Factors Affecting Fertility*).

Low birth rates are associated with low death rates (fig. 4). Low birth rates reflect the use of foresight and prudence in limiting conception rate to the parents' ability to rear their children in harmony with their child-welfare ideals, and intelligent adjustment of a birth rate pattern evolved for preliterate, prescientific, conditions when life expectancy was 14 years, to a scientific era in which the life expectancy exceeds 70 years. It appears that such ideals are lacking among the peoples, half of whose children die before reaching age 13 (Ref. 18 and fig. 4) mostly as result of inadequate food supply.

FOOD SUPPLY

1. Limiting factors. While awaiting the development of effective methods for retarding the rate of world-population growth, we are impelled

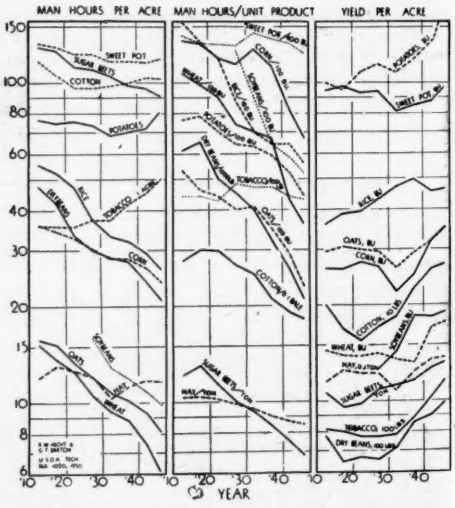


Fig. 5. IMPACT of developing science and technology on efficiency and level of production of food of plant origin (plotted from indicated source). Most agriculturists believe accelerating crop production accelerates impoverishment of soil when food is shipped to the cities and excreta washed to the ocean or incinerated.

by our ethics—and possibly by our diplomacy—to do our best toward increasing food production to feed the hungry, although this is likely to aggravate the later future by further population increase. The rate of food production can undoubtedly be increased by various technological and educational aids and developments (figs. 5 and 6). In this sense there is much good in the FAO and Point Four programs.

There are, however, definite limits to these developments. Just as limiting factors prevent

conversion of 100 per cent of the coal energy into useful work of a steam engine, so limiting factors prevent cultivation of 100 per cent of the earth's surface, utilization of 100 per cent of the solar energy poured on the earth, greater exploitation of yeast and chlorella growth potentialities. Application of fertilizers increases crop yield, but their supply, especially of potassium, phosphorus, and sulphur, is limited. Certain soils are not cultivated either because they are not accessible or because of their relative sterility. The present productive ceilings can be raised, but subject to

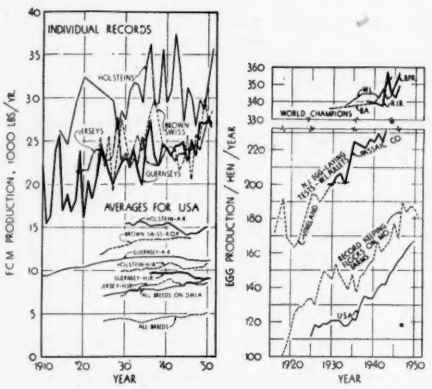


Fig. 6. Impact of developing agricultural science and technology on milk and egg production in USA (FCM = milk in terms of 4% fat). Individual records represent maximal production records. Milk production data furnished by respective cattle clubs; D.H.I.A. data of all breeds supplied by J. F. Kendrick, U. S. Department of Agriculture; average egg production, courtesy H. L. Kempster, U. S. Department of Agriculture; Missouri farm flocks, *Poultry Record Summary* 6: No. 10, University of Missouri Agricultural Extension Service; New Jersey tests, C. S. Platt, *N. J. Agric. Exper. Stn. Bull.* 720; champion data, P. H. C. du Plessis, *World's Poultry Sci. J.* 5: 150, 1949.

the 'law of diminishing increments' (20); and while helpful locally such elevation will be insignificant in comparison to the increasing world food needs.[3]

Such 'little things' as carbon dioxide concentration in the atmosphere, rainfall, temperature,[3] latitude, soil fertility, wind, topography, etc., limit greater utilization of soil and sun energy. It is theoretically possible to increase the carbon dioxide concentration in the atmosphere, to maintain plants under optimal temperature conditions, to irrigate the deserts, to drain the swamps, to pulverize and level rocky mountains, to produce yeast and chlorella on a world-wide scale of gigantic dimensions, to increase the skill, intelligence, and ethical standards of food producers and consumers, but is it practicable?

2. Food of plant versus animal origin and the efficiency of livestock production. There are good reasons for livestock production, as follows:

a) *Ruminants* harvest the vegetation on marginal land not suitable for cultivation, and consume roughages (hay, silage, straw) and other high-cellulose feeds (bran, chaff) not suitable as food for man. Ruminants also harvest and consume cultivated grasses, which under the best conditions yield 1.5 times as much available energy per acre as potatoes and 3 times as much energy as grain. Such grasses may also yield 2 times as much protein as grain or 3 times as much as potatoes (21). Moreover, ruminants synthesize (by way of their rumen microorganisms) nutritionally essential amino acids and B vitamins. They are not only converters and refiners, but also creators of new nutrients, and so contribute mightily towards maintaining a high nutritional level.

b) *Non-ruminants* (pigs and chickens) which, unlike ruminants, resemble man in their dietary needs, are useful as refiners of food of plant origin, and of various wastes, such as garbage, that have nutritional value yet not edible by man, into highly prized human foods.[4] These animals are also useful scavengers.

c) In periods of abundance farm animals are fed surplus grain; in periods of scarcity they are liquidated, yielding their bodies and releasing the grain which they would have consumed. They thus serve as 'shock absorbers' saving refined food for man against the lean years. Overpopulated regions not having such reserves are 'adjusted' to the decreased food supply by the four horsemen: war, pestilence, famine, and death (22).

d) Foods of animal origin—milk, meat, egg—supply essential amino acids (particularly lysine, methionine, tryptophane) and B-vitamins (particularly B_{12}) in which foods of plant origin are deficient.

The efficiency of conversion of food of plant origin to that of animal is, however, relatively low. Many more persons can be supported from an acre of arable land when they consume the plants directly then after converting to food of animal origin. Hence, poverty is associated with

[3] Climatic changes may render arctic lands suitable for food production. This will be an *arithmetical* addition, quickly overtaken by *exponential* population increase.

[4] "With a garden and a cow
Using our wastes with a sow
Small families could get along somehow."
M. E. Muhrer

a diet low in food of animal origin (fig. 7), with consequent higher child mortality.

Experimental data on rats (23) and observations on man (24) indicate that mature males—or non-reproducing females—can maintain good nutrition indefinitely on an all-vegetable diet containing such foods as soybean meal. They are, however, inadequate for reproducing females and growing animals, perhaps because such foods—even soybean—lack vitamin B_{12}.

however, inactivated by cooking. Although soybean diets show a high protein-efficiency ratio body weight increase/protein ingested in *short* feed experiments, there is no full agreement on the results when fed to children over *long* periods (25, 26). Fair nutrition has been achieved in China on all-plant diets using the soybean as base, perhaps because of the B_{12} in their sauces and other fermented products. Pigs, chickens, and rats have been reared on all-plant rations

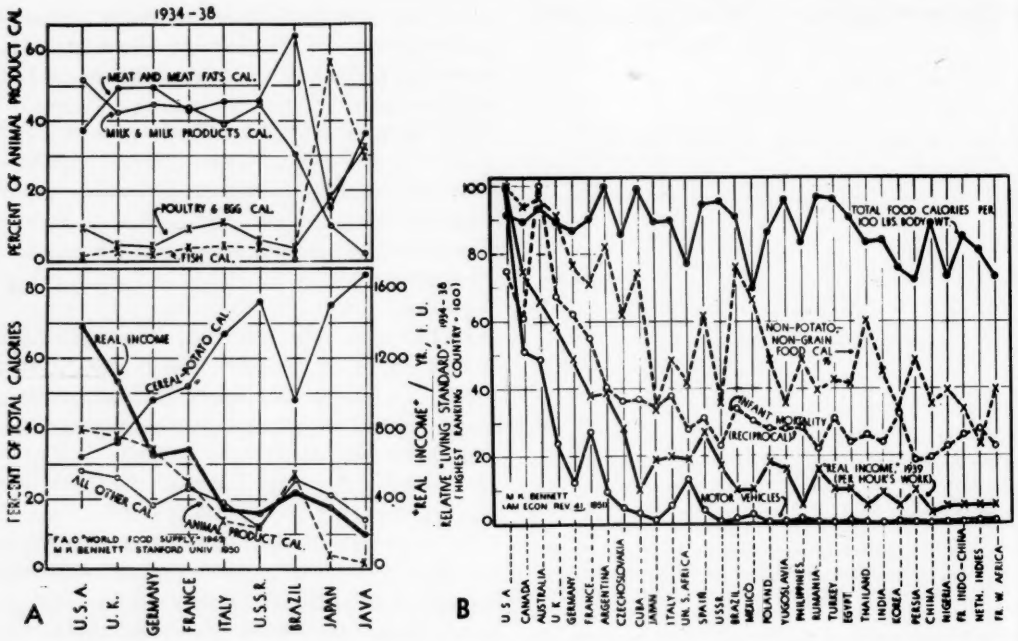


Fig. 7. (*Left*) ANIMAL PRODUCTS in national diets. Note parallelism between 'real income' and consumption of food of animal origin. (Plotted from paper of same title by M. K. BENNETT, Food Research Inst., Stanford University; transmitted to author by JOSEPH S. DAVIS, Director, Stanford Research Institute. (*Right*) EXTENSION of figure to more nations, showing parallelism of nature of diet with the reciprocal of infant mortality, with 'real income' and with 'living standard' as indicated by use of motor vehicles (from data published by M. K. BENNETT; reference on chart).

Children (Bantu in Africa) fed after weaning on a diet exclusively of plant origin (mealie-meal porridge, Kaffir corn, green vegetables) develop 'malignant malnutrition' or 'kwashiorkor.' This is characterized by anorexia, anemia, diarrhea, enlarged fatty liver, edema, abnormally low serum protein, atrophy. Including soybean in the diet did not prevent or cure this condition. Milk is, so far, the only known prophylactic and cure for this malnutrition (25).

The soybean comes nearest to having the limiting amino acids found in food of animal origin. Its lysine and methionine-cystine levels are of the same order as in meat meal (26). Soybeans contain a trypsin inhibitor which is, containing soybean fortified with vitamin B_{12} and calcium (23).

Summarizing, cereals and potatoes are the most economical dietary energy sources for man. They are, however, deficient in calcium; in vitamins B_{12} and A, in lysine, tryptophane, and methionine. Meat, eggs, and milk are rich in these nutrients; hence, optimal nutrition involves the supplementation of the cereal-potato diet with food of animal origin. Such supplementation appears to be indispensable for children and reproducing females. All-plant diets containing soybean, or Puerto Rican chick or pigeon peas (27), or tender cereal grass (28) could, perhaps, be made adequate even for children if supplemented

with vitamin B_{12}. Now that the essential vitamins and amino acids that are deficient in food of plant origin are factory produced, cereal-potato diets may, perhaps, be made nutritionally complete by fortification with these nutrients. This way the great losses during conversion of plants to food of animal origin would be avoided, and the population-supporting capacity of the land would be greatly increased.

The efficiency of conversion of plants to food of animal origin varies with the animal and its product, and with the nature of the feed. With respect to the digestible feed consumed, the immediate energetic conversion efficiency for superior animals under good management on excellent feed is about 33 per cent for milk production; 17 per cent for egg production; 35 per cent for pork production (including the lard); and 5 to 33 per cent for other kinds of meat production, depending on age (29). The *protein* conversion efficiency is about 33 per cent for milk production and egg production, and about 25 per cent for pork production. The conversion efficiency is, of course, very much less, about half—3 to 17 per cent—if it is computed with respect to the *gross or total dietary energy* (including the non-digestible part). There is a large literature on the efficiency problem but it is confused by the use of different reference bases—digestible energy, gross energy, metabolizable energy, net energy—each of which gives a different efficiency value (30).

The social economy of conversion of plant feed to meat, milk, or egg, is not necessarily correlated with machine efficiency. For instance, the conversion of western range pasturage to meat is not efficient, but it is economical, because this land cannot be used for grain or potato production. Likewise, the use of swine and poultry as scavengers for producing good food from humanly useless offal may not be efficient but it is extremely economical. The production of milk on excellent grassland is highly efficient—a superior dairy cow on first grade ryegrass pasture may produce in milk energy over 50 per cent of the energy value of cereal produced on this land—but not necessarily socially economical during a period of serious food-calorie shortage, since the need for calories is primary for a starving community. This explains how consumption of food of animal origin declines rapidly with decline in fertile land per person, and with decline in real income (fig. 5). For instance, 25–35 per cent of the dietary calories is from food of animal origin in prosperous USA, Australia. and New Zealand; 15 per cent in less prosperous Europe and in most of South America; and only about 3 per cent in economically depressed Asia. The USA feeds 60 per cent of its grain to livestock; Asia feeds none of her grain to livestock (22).

Milk appears to be the most economical supplement to food of plant origin, not only because of efficiency and economy of its production, but also because it contains at a uniform level—at virtually no cost—the limiting amino acids and B vitamins, including B_{12}. It is the richest dietary source of calcium. Its slightly soluble carbohydrate, lactose, passes into the intestine where it becomes a favorable medium for vitamin B synthesis and promotes calcium and phosphorous assimilation.

Unfortunately, milk production declines rapidly with increasing environmental temperature because cattle (unlike man) cannot increase the evaporative cooling with increasing environmental temperature above about 80°F (27°C), and so develop fever. The milk yield is, therefore, lowest where it is most needed as supplement to diets of plant origin, namely in the overcrowded tropical and semitropical regions. This problem is at present under intensive investigation (31).

ATTITUDES AND OPPOSITION

The attitudes of man towards the universe in general and to population growth in particular reflect his culture. Much of the history of science is history of cultural lags and struggles between old and new ways of thinking about nature and human nature (3). Our present demographic crises, with their associated social disturbances, resulted from such cultural lags, from an archaic combination of the application of modern scientific medicine for reducing child mortality with a prescientific rate of child birth.

The current life expectancy at birth at least in the USA, is about 70 years. The reproduction rate in man—its biological mechanisms and supporting traditions (including religions)—were evolved to maintain the species in an era when life expectancy was 14 years. These traditions make it praiseworthy to apply the newest methods for prolonging life but not for reducing the birth rate ("because frustation of conception is wrong" and "whoever interferes with the God-established process of perpetuating human life interferes with human life itself"). The result is an explosive human population upsurge with rise in modern medicine analogous to a rabbit

population upsurge on removal of his natural enemies.

The traditionally-bound regions as regards birth rate include the Far and Near East, most of the Eastern European countries, and certain South American regions, including our own Puerto Rico, which furnishes a well documented case history (32). A Professor of Biochemistry in Puerto Rico writes (33):

> The main trouble in P. R. is that while the death rate declines, birth rate remains fairly constant. Regardless of religious affiliation, the women of the upper economic classes have small, while the poor have large, families. The latter are demanding the information and facilities which the former employ in limiting the size of their families.

and a Congregational Medical Missionary in Puerto Rico writes (34):

> How can we suppose that God . . . could want children to be born into homes where there is no possibility of health. It seems that God would want man to use his intellect, not his emotions and sentiments, to work out a solution of this problem. Our Christian ethic raises its head in protest against the starvation. It is unthinkable that each year an additional 30 or 40 thousand could be accommodated . . . yet this is the rate at which our population (in Puerto Rico) is increasing . . . We let children come in swarms . . . even though filth, ignorance, insecurity, will be all the days of their life . . . even though in some areas nearly half of them are doomed to die, and those who survive exist in sordidness and shame.

The above 1952 comment is similar to one by Malthus (14), an Episcopalian divine of 150 years ago:

> The Creator cannot have commanded increase without relation to means of support . . . Our duty is to seek the greatest sum of human happiness . . . A common man who has read his Bible must be convinced that a command given to a rational being by a merciful God cannot be interpreted as to produce only disease and death instead of multiplication.

The following is from an Indian nutritionist (35):

> The Father of our Nation, Mahatma Gandhi, was alarmed with the growth of population in India and he visualized starvation and semistarvation of the mass of people. He suggested moral restraint, but it did not work. Sex is the only recreation in their life . . . Their philosophy is that one must take the gift of God and not grumble.

The Oriental attitude is said to consider earthly life as a transient phenomenon, and there is no point in prolonging it. The present government in India, however, is encouraging procreative restriction. It has requested the UN World Heath Organization for help (35) for a program of birth control. Family planning centers are being established in Delhi, Mysore, Madras, West Bengal, Poona, and elsewhere; research is being conducted on indigenous and recent contraceptive methods, and villagers seek, and are given, information. In this case leaders in India seem to be more concerned in alleviating starvation among their poor than those in Puerto Rico, who would rather see, as Dr. John Smith (34) phrased it, "half of the children doomed to die," than permit physicians to offer help for what appears as reasonable family planning that would avoid unnecessary wastage of human life.

Some categories of opposition towards family limitation are based on curious considerations. For instance, the latest FAO reports (2) show that it is apparently impossible to keep up the food production with population increase in free-reproducing societies in which the major human enemy, communicable disease, is controlled by governmental agencies; they indeed substantiate the operation of the old 'Malthusian Principle,' and refute the notion of 'permanent economy of abundance.'

Yet Josue de Castro, Chairman of the FAO's policy-making Executive Council, hopefully—and angrily—writes (36), as if he never heard of the FAO's reports:

> The empty sleeve of the Malthusian scarecrow has flapped in the winds of prejudice for a century and a half, but science and history have finally shown that no one need take it seriously.
>
> Starvation actually causes overpopulation . . . protein deficiency leads to cirrhosis of the liver which causes release of more estrogen . . . increases sexual appetite . . . and reproductive capacity.
>
> Chlorella, an alga, which can be grown commercially in tanks and fed with minerals, is rich in proteins. A single installation could raise enough of it to supply the proteins needed by three million people . . . equivalent to the production of 150,000 acres of good arable land.
>
> At the present time only about 10 per cent of the earth's arable land is being cultivated.

These statements are contrary to the facts. No one has demonstrated that the starvation or malnutrition in humans increases their sexual appetite or fertility (37), or that a nourishing diet

including meat, milk, and eggs is contraceptive; no commercial installation of algae has been made; the world's accessible arable land is being cultivated as testified by John Boyd Orr, first Secretary-General of FAO, who said that "the last new land which could easily be brought under the plow was broken during World War I. . . . We have exploited the last of the world's virgin soil."

De Castro's book (38) presents a table showing that Formosa, subsisting on 4.7 gm. animal protein a day, has a high birth rate whereas Sweden, enjoying 62.6 gm. animal protein, has a low birth rate. The author concludes[5] therefrom that a poor diet causes a high birth rate and a good diet a low birth rate. This is a remarkable example of confusion between cause and effect. Is it not more logical to assume that the less tradition-bound Swedish population plans to have no more children than can be reared in harmony with their high living standard? and that this planning enables them to have 63 gm. animal protein?

GENERAL SUMMARY

Increasing food production is meaningful only when expressed in relation to increasing population, as ratio food/population. The FAO of the UN reports (2) that this ratio is steadily declining; that "production is . . . failing to keep pace with population growth . . . which means . . . more funeral processions . . . rising tide of unrest and revolution and war."

This lag in food production behind population growth, in spite of Point Four, Colombo Plan, Marshall Plan, gifts and loans, and related heroic measures, is caused by a historically unprecedented world population upsurge (39) currently at 1.2 per cent per year or 260 million per decade (table 1). A simple computation shows that such a population growth rate would soon saturate the surface of the earth.

This historically unprecedented rate of population growth is causally related to an archaic combination of modern scientific public health control (life expectancy over 70 years in the USA) with an ancient reproductive pattern evolved for a life expectancy of about 14 years. The present population upsurge associated with decline of fatal infections is analogous to an upsurge of subhuman population growth on the removal of natural enemies. Such a combination of declining death rate with constant birth rate can not long coexist. Either the birth or conception rate will be reduced prophylactically to approximate the death rate; or else the population will be scaled down to the available food supply by the ancient population-adjusting methods: hunger, disease, violence, and death. These latter patterns are now in operation in many parts of the world where half of the children die before reaching the age 13 (18)—a sacrifice to cultural lag.

The confusing details of the population-food problem result in confusing opinions about its solution. The general fact, however, is clear. The human population has the capacity and 'will' for indefinite exponential growth (in compound-interest fashion) whereas the population-supporting capacity of the earth is limited. Hence, the freedom from want ideal cannot be made a world-wide reality unless the decline in birth rate is adjusted to approximate the decline in death rate. The task of adjusting the birth or conception rate to the death rate is *technically* simple as illustrated by the virtually stationary populations (fig. 1) in Ireland, France, Belgium, and can be further simplified (40). This task is, however, *culturally* formidable in most of the world because of the difficulty in enlisting the cooperation of the leaders of the traditional mass mind.

The opposition by certain religious groups to rational conception control is puzzling (3). This is particularly true since the broad, often poetic, language and spirit of the sacred writings of most of the old traditional religions encourages compassionate, humane, interpretations. For instance, we may interpret the conception-control problem with Jesus who said: "The Sabbath was made for man and not man for the Sabbath" (Mark II, 27); that man was endowed with judgment so that he may use it with compassionate discretion for solving problems developed by changing conditions. Medical knowledge and technology for family planning are already available to the literate well-to-do but not to the poor and ignorant. The latter are offered pious but unrealistic, unacceptable, advice. Such disparity in

[5] This is based on the Thomas Doubleday "Law" given in his "The Law of Population" etc. (1853) in answer to Malthus: "Whenever a species or genus is indangered, a corresponding effort is invariably made by nature for its preservation and continuance, by an increase of fecundity or fertility, and that this especially takes place whenever such danger arises from a diminution of proper nourishment". (Courtesy Robert C. Cook, acting Director Population Reference Bureau, March 24, 1952.)

medical information is unethical, undemocratic and socially harmful, because it is promoting—indirectly but effectively—a differentially higher birth rate among the poor and ignorant.

Considerable attention was also given to the food production factor in the food/population ratio. The population-supporting capacity of the earth can be greatly increased but at progressively increasing social costs.

Grateful acknowledgments are made to my UM associates and to many correspondents, particularly J. S. Davis and H. C. Farnsworth (Food Research Institute, Stanford University) for valuable help.

REFERENCES

1. Carlson, A. J. *Proc. Inst. Med. Chicago* 19: No. 1, 1952.
2. Dodd, N. E. *The Work of the FAO*. Report of Director General, Food & Agricultural Organization of U.N., Rome, 1951; also *Item III, Provisional Agenda, Conference of FAO, Nov. 19–Dec. 7, 1951*. FAO/51/10/2780.
3. Russell, B. *New Hopes for a Changing World*. New York: Simon & Shuster, 1951; also Frank, L. *Nature and Human Nature*. New Brunswick: Rutgers Univ. Press, 1951.
4. Carr-Saunders, A. M. *The Population Problem, a Study in Human Evolution; Population; World Population: Past Growth and Present Trends*. New York: Oxford Univ. Press, 1922, 1925, 1936. See also Buliard, R. *Inuk*. New York: Farrar, Straus & Young, 1951.
5. Richards, A. I. *Land, Labor and Diet in Northern Rhodesia*. New York: Oxford Univ. Press, 1939.
6. Dodd, Norris E. *Chronica Botanica* 11: 207, 1948.
7. Myrdal, A., K. Davis, J. J. Spenger, H. A. Spoehr, A. J. Stann, L. A. Walford and M. T. Jenkins. *Proc. Am. Phil. Soc.* 95: 1, 1951.
8. Fawke, V. C. *Can. J. Econ. & Pol. Sc.* 17: 501, 1951.
9. Hobson, J. A. *Imperialism: A Study*. London: Constable, 1905.
10. Mukerjee, R. *Races, Lands and Food*. New York: Dryden Press, 1948.
11. Thompson, W. S. *Ann. Am. Acad. Sci.* 249, 111, 1947.
12. Vogt, W. The Road to Survival. New York: 1948.
13. Franklin, Benjamin. *Works, Vol. 2* (Sparks, *Ed.*) Boston: Hilliard, 1840.
14. Malthus, T. R. *Essay on the Principle of Population. 1st ed.* 1803; *9th ed.* 1826. London: MacMillan.
15. Darwin, Charles. *Origin of Species*. London: J. Murray, 1859.
16. Pearl, Raymond. *Biology of Population Growth*. New York: Knopf, 1925; *The Natural History of Population*. New York: Oxford Univ. Press, 1939. See also numerous other books and articles by Pearl and associates.
17. Brody, S. *Bioenergetics and Growth*. New York: Reinhold, 1945, pp. 490–92, 501, 520, 544–46.
18. Balfour, M. C., R. F. Evans, F. W. Notestein and I. B. Taeuber. *Public Health and Demography in the Far East*. New York: Rockefeller Foundation, 1950; also Davis, Kingsley. *The Population of India and Pakistan*. Princeton: Princeton Univ. Press, 1951.
19. *Demogrpahic Yearbooks, 1949–50, 1950–51*. New York: UN Statistical Office, Depts. of Economic and Social Affairs.
20. Brody, S. *Bioenergetics and Growth*. New York: Reinhold, 1945, pp. 1 and 76.
21. Boyd, D. A. *Brit. J. Nutr.* 5: 255, 1951.
22. Pearson, F. A. and D. Paarlberg. *Starvation: Truths, Half Truths, Untruths*. Ithaca: Cornell Univ., 1946.
23. Hogan, A. G. Personal communication. See also Hogan *et al. J. Nutr.* 40: 243, 1950; *Proc. Soc. Exper. Biol. & Med.* 76: 349, 1951; *Nutr. Abstr.* 19: 781, 1950.
24. Mirone, L. *Science* 111: 673, 1950.
25. Wills, L. *Brit. J. Nutr.* 5: 265, 1951. See also Dean, R. F. A. *Id.* p. 269.
26. Carpenter, K. J. *Brit. J. Nutr.* 5: 243, 1951; also Chick, H. *Id.* p. 261.
27. Roberts, L. J. and R. L. Stefani. *Patterns of Living in Puerto Rico*. Puerto Rico: Univ. of Puerto Rico Press, 1949.
28. Brody, S. *Bioenergetics and Growth*. New York: Reinhold, 1945, pp. 749, 789.
29. Brody, S. *Bioenergetics and Growth*. New York: Reinhold, 1945, Ch. 3, also pp. 827–47, 880–89, 900–14.
30. Brody, S. *Bioenergetics and Growth*. New York: Reinhold, 1945, pp. 5–6, 24–32, 54–8. Cf. Forbes and Swift. *J. Dairy Sc.* 8: 15, 1925; Forbes and Voris. *J. Nutr.* 5: 395, 1932; Maynard, L. A. *J. Nutr.* 32: 345, 1946.
31. *Research Bulletins, Environmental Physiology Series*. Columbia: Univ. of Missouri Agricultural Experiment Station.
32. Janer, J. L. *Human Biol.* 17: 267, 1945.
33. Goettsch, Marianne. Personal communication.
34. Smith, John. Personal communication.
35. Mullick, D. N. Personal communication.

36. De Castro, J. *Nation* Feb. 16, 1952, p. 156; *Colliers*, Jan. 19, 1952.
37. Carlson, A. J. and F. Hoelzel. *J. Gerontol.* 6: 169, 1951; also *Overnutrition: Its Causes and Consequences.* Springfield: Thomas, 1950; Evans, H. M. and M. M. Nelson. *Federation Proc.* 11: 45, 1952. See also Benedict, F. G. *et al. A Study of Prolonged Fasting.* Carnegie Inst. of Wash. ington, 1915; Keys, A. *et al. Human Starvation.* Minneapolis: Univ.of Minnesota Press, 1950.
38. De Castro, J. *The Geography of Hunger.* Boston: Little, Brown, 1952, p. 72.
39. Davis, J. S. *The Population Upsurge in the U. S.* San Francisco: Stanford Univ. Press, 1949; also *J. Farm Econ.* 31: 765, 1949.
40. Adams, R. *Science* 115: 157, 1952.

AMERICAN PHYSIOLOGICAL SOCIETY

Symposium on Secretion of Electrolytes[1]

Chairman: FRANKLIN HOLLANDER

RENAL SECRETION OF POTASSIUM AND HYDROGEN IONS

ROBERT W. BERLINER

From the Section on Kidney and Electrolyte Metabolism, National Heart Institute, National Institutes of Health, Public Health Service, Federal Security Agency, Bethesda, Maryland

THE term secretion is generally used in renal physiology to signify transport from tubule cell to tubule lumen. This is a convenient designation to distinguish transport in this direction from that which effects removal of substances from the tubule lumen, the latter being referred to as reabsorption. In the broader sense of electrolyte transport to which this symposium is devoted, both are frequently active secretory processes, effecting net transfers of substances against gradients of chemical potential. To avoid confusion, however, it will be convenient for our present purposes to designate as secretion only that which is directed from cell to lumen while referring to the two processes, secretion and reabsorption, collectively, by the more general term, transport.

Potassium transport is complicated by the coexistence of both secretory and reabsorptive processes. This may seem somewhat paradoxical when it is recognized that the amount of potassium usually excreted is only some 10–15 per cent of the amount made available for excretion by the process of glomerular filtration. Except in unusual circumstances, simple failure to reabsorb the amount filtered would be quantitatively adequate to provide the necessary excretory capacity. However, extensive and conclusive evidence now indicates that the tubules do have the capacity to secrete as well as to reabsorb potassium, and, in fact, there are some reasons for suggesting that the secretory process contributes significantly to normal potassium excretion even though the total excreted is far below the amount filtered.

Although the possibility of potassium secretion by the renal tubules had been suggested by several earlier clinical observations (1, 2), these findings had been all but completely overlooked by those interested in renal physiology. Interest in the possible role of this mechanism in potassium excretion has been aroused more recently by the definitive experimental demonstration that potassium excretion can regularly be made to exceed the limits set by the amount in the glomerular filtrate, particularly by the administration of large amounts of potassium salts (3), but also by osmotic diuresis (4) and by the administration of hypertonic solutions of sodium bicarbonate (5).

There is practically no information concerning potassium *reabsorption* by the renal tubule. Beyond mentioning the suspicion that it occurs early in the passage of fluid down the tubule and the probability that under most conditions the major portion of the filtered potassium is so transported, very little can be said about the nature of the process.

Since the thesis to be developed depends largely upon certain experiments which suggest the nature of the process involved in the secretion of potassium, its relationship to the secretion of hydrogen ions and to the transport of sodium ions, attention will be directed to these studies. Time will not permit consideration of a number of important factors which have been found to influence the rate of potassium excretion.

The possibility that the secretion of potassium might be an exchange of potassium ions for sodium ions was suggested by the reciprocal relationship between the excretion of these ions under certain conditions. Such a cation exchange mechanism had already been suggested to explain the secretion of hydrogen ions by the renal tubules (6). This hypothesis seemed susceptible to ex-

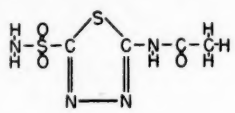
[1] New York City, April 16, 1952.

perimental verification as follows: If, during a period of net potassium secretion, one could show that the urine excreted contained more secreted potassium than anion which might have been secreted with it, then one must conclude that potassium entered by some process which involved no net secretion of anion. In the case of an anion which is consistently excreted at the rate at which it is filtered, it is reasonable to assume that it is neither secreted nor reabsorbed. Though such an assumption has since been shown to be invalid in the case of thiosulfate (7), ferrocyanide in the dog (8, 9), still appears to fulfill the necessary requirements. Accordingly, Drs. Kennedy, Hilton, and I examined the electrolyte content of the urine of dogs receiving infusions of large amounts of potassium ferrocyanide (10). Under these conditions, the amounts of potassium excreted were often 3–400 μEq/min. greater than the amounts filtered. That is, at least 3–400 μEq/min. of potassium were secreted. At the same time, the anion in the urine was largely ferrocyanide. The sum of chloride, bicarbonate, sulfate, and phosphate fell short of covering the secreted potassium by something between 150 and 300 μEq/min. While there is no reason for believing any of these latter are secreted, even conceding this possibility would not account for the secreted potassium. While the possibility that some potassium salt might be secreted and the corresponding sodium salt reabsorbed further down the tubule could not be entirely excluded, it seemed more probable that the mechanism is actually a direct exchange of potassium ions for sodium ions, particularly when account is taken of the fact that the secretion of potassium probably occurs in the more distal portion of the tubule (10).

This mechanism parallels in its essentials that described for the secretion of hydrogen ion by Pitts and his associates (6, 11). This was reviewed before a Joint Session of the Federation several years ago (12), and only certain essential features of their work need be considered here. It was shown that secretion of hydrogen ions was required to account for the high rate of excretion of titratable acid achieved by the administration of large amounts of buffer to acidotic individuals. The absence, from urine obtained under these conditions, of anions which were thought to be secreted led to the conclusion that the hydrogen was secreted by cation exchange for sodium. It was shown that sulfanilamide in massive dosage had an inhibitory effect on acid excretion, and it was concluded that carbonic anhydrase played a role in the process by which hydrogen ions were made available for secretion. This mechanism for exchanging hydrogen ions for sodium ions, partially dependent upon carbonic anhydrase, was viewed as localized to the distal convoluted tubule and responsible for the excretion of titratable acid and the reabsorption of that portion of bicarbonate which escaped reabsorption in the proximal segment. The reabsoption of bicarbonate in the proximal segment was interpreted as a different type of process, apparently involving transport of both cations and bicarbonate ions. The latter process was considered to be responsible for some 80 per cent of bicarbonate reabsorption when this was occurring at its maximum rate, and to effect complete reabsorption of bicarbonate when the concentration of this ion in the plasma fell below about 20 mEq/l.

Certain observations which indicate that the secretion of hydrogen ion and potassium ion are related may be considered against this background of the overall similarities of the two processes. These observations fall into two groups —one relating acid-base balance to muscle potassium content, and one specifically relating body potassium content and excretion with urinary acidification. While complete data are not available to indicate that these two groups of observations are different aspects of a single phenomenon, the suspicion that this is the case is strong. The relationship between muscle potassium content and acid-base balance was first described in detail by Darrow and his associates (13, 14). It was found that, in general, animals depleted of potassium became alkalotic while animals made alkalotic became potassium depleted. It was also noted that in acidotic animals, the muscle potassium concentration increased. It was suspected that normal renal function was required for the development of these related changes, but detailed data showing the role of the kidney in this process were not available. Actually, it has been known for some time that the administration of neutral potassium salts causes the urine pH to rise (10, 15, 16). Conversely, the excretion of potassium is increased by alkalosis (5). In a few instances in which studies have been made, it has been found that alkalotic, potassium-depleted individuals excreted a urine which was inappropriately acid and became alkaline only when potassium salts were administered (17, 18). These observations remained unexplained until the relationships were made even more apparent by the effects of certain carbonic anhydrase inhibitors.

Because of the demonstrated effect of sulfanilamide on urine acidification (6), interest in the possible use of carbonic anhydrase inhibitors as diuretics was aroused (19). Among several very powerful inhibitors developed by American Cyanamid (20), one, which is shown in figure 1, is being subjected to extensive study. This compound, 2-acetylamino-1,3,4-thiadiazole-5-sulfonamide has been used in a series of studies by Drs. Orloff, Kennedy, and myself (21). It has a very striking effect on urine acidification; when it is injected intravenously in acidotic dogs, the urine becomes alkaline almost instantaneously. Two examples are shown in tables 1 and 2. In the control periods, these mildly acidotic dogs excreted acid urine containing appreciable amounts of titratable acid. Following the injection of ✱6063, the urine became distinctly alkaline, all titratable acid disappeared and was replaced by large amounts of bicarbonate. As would be expected,



2-Acetylamino-1,3,4 thiadiazole-5-sulfonamide

Fig. 1

since a process which is presumably one of sodium reabsorption by exchange for hydrogen ions had been inhibited, there was a very striking increase in sodium excretion. Not quite so expected was the marked increase in the excretion of potassium. While smaller than the increment of sodium on an absolute basis, the increases in potassium excretion were much larger relative to the amounts of each cation filtered. In a number of such experiments, the increment in potassium excretion was larger than the amount filtered. An increase of this magnitude could occur only if there was some augmentation in secretion of potassium. That actually all or most of the increment is attributable to secretion can be deduced from the effect of mercurial diuretics on the augmented excretion produced by ✱6063. This is shown in tables 3 and 4. Each figure in these tables represents the mean of three consecutive periods and indicates successively control, effect of salyrgan or ✱6063 alone, combined effect of salyrgan and ✱6063, and finally, after the effect of salyrgan is abolished by BAL, the action of ✱6063 alone. In table 3, the effect of the mercurial is indicated by the marked increase in the excretion of chloride and sodium. The superimposed effect of ✱6063 is indicated by the increase in bicarbonate excretion and the further rise in sodium excretion. Note, however, that during mercurial action there is essentially no effect of ✱6063 on potassium excretion. Finally, when the mercurial effect is removed, as indicated, by the return of chloride excretion to control levels, potassium excretion goes up to the level usually observed after ✱6063.

TABLE 1. ELECTROLYTE EXCRETION BEFORE AND AFTER ADMINISTRATION OF ✱6063 TO MILDLY ACIDOTIC DOGS (*Dog H, 26 kg.*)

TIME min.	URINE FLOW ml./min.	URINE pH	EXCR. Titr. Acid μEq./min.	EXCR. HCO_3^- μEq./min.	EXCR. Na^+ μEq./min.	EXCR K^+ μEq./min.
-77	Start infusion of Creatinine 25 mg./min.; Sodium Phosphate (pH 5.8) 200 μM./min.; Sodium Chloride 1 mM./min.; Water 8 mL./min.					
-57	Priming Infusion: Creatinine 2.0 gm.; Hydrochloric Acid 40 mEq.					
0-21	0.81	5.58	105	0.2	85	68
21-43	0.77	5.41	154	0.2	85	61
43-64	0.94	5.38	164	0.3	116	52
73-79	Infusion ✱6063 (10 mg./kg.)					
87-102	8.06	7.82	0	920.0	1654	264
102-117	6.21	7.80	0	750.0	1315	248
117-133	4.17	7.73	0	456.0	935	184

TABLE 2. ELECTROLYTE EXCRETION BEFORE AND AFTER ADMINISTRATION OF ✱6063 TO MILDLY ACIDOTIC DOGS (*Dog M, 15 kg.*)

Time min.	Urine Flow ml/mn.	Urine pH	Excreted Titr Acid μEq/min.	Excreted HCO_3^- μEq/min.	Excreted Na^+ μEq/min.	Excreted K^+ μEq/min.
-44	Infusion: Sodium Phosphate (pH 5.8) 198 μM/min.; Creatinine 16 mg./min.; Water 1.8 ml./min.					
-30	Prime: 1.5 gm. Creatinine					
0-24	0.68	6.40	56	2.0	69	79
24-44	1.18	6.48	65	3.9	126	85
44-56	1.07	6.28	80	2.1	113	82
63	✱6063 i.v. (20 mg./Kg.)					
69	Infusion ✱6063 (40 mg./Kg./hr.)					
80-96	3.90	7.70	0	362	566	173
96-111	3.38	7.73	0	358	537	164
111-124	3.17	7.74	0	380	520	159
124-136	3.31	7.72	0	338	555	179

In the experiment in table 4, the drugs were administered in reverse order. Potassium excretion was restored to control levels when the mercurial effect was superimposed on that of ✱6063 and rose again after BAL. Failure of chloride and sodium excretion to be reduced by BAL in this instance is unexplained. It is apparent that the

enhancement of potassium excretion due to ✱6063 can be prevented, or abolished, by the action of mercurial diuretics. Such a combination of events could be produced only if the effect of ✱6063 led to an increase in the secretion of potassium while the mercurial inhibited potassium secretion (21). Since the potassium is, then, added to the urine in exchange for sodium already present, we cannot conclude that the potassium excretion has been forced, for the maintenance of electroneutrality, by the bicarbonate whose reabsorption has been inhibited. Any explanation must take into account the fact that there has been an increase in the secretory process. It is also worthy of consideration that this is not a phenomenon produced only by carbonic anhydrase inhibitors. Figure 2 indicates the similarity in the effects of several agents with what are presumably fundamentally different mechanisms of action. In this figure potassium excretion and urine pH are plotted for successive urine collection periods before and after administration of the various indicated substances. To emphasize the inverse relationship of acidification and potassium excretion, the pH is plotted from above down. The relationships are roughly similar for all of these: ✱6063 which is a carbonic anhydrase inhibitor, maleate which has no activity against carbonic anhydrase (22) but inhibits urine acidification (23), lithium which had previously been shown to enhance the secretion of potassium (24) and which has now been found to inhibit acidification of the urine (25) but not carbonic anhydrase (24), and finally potassium itself. These observations, along with those already referred to relating urine acidification and potassium excretion, seem most logically explained by the hypothesis that in the secretion of potassium and hydrogen ions, both species enter into reversible combination with some common component so that the presence of one tends to exclude the other. Then, other things being equal, it would be expected that an increase in potassium would reduce hydrogen ion secretion, a decrease in potassium would favor acid secretion, a decrease in hydrogen ion concentration would favor potassium excretion, etc. Time will not permit a discussion of the extent to which factors other than this competition *alone* might contribute to the development of some of the clinical and ex-

TABLE 3. ELECTROLYTE EXCRETION IN CONTROL PERIODS AND AFTER SUCCESSIVE ADMINISTRATION OF SALYRGAN, ✱6063 AND BAL (*Dog V*)

Time	C_{cr}	Excretion of			
		Cl^-	HCO_3^-	Na^+	K^+
min	ml/min	μEq/min	μEq/min	μEq/min	μEq/min
-69	Start infusion of creatinine, phosphate (pH 7.4), saline.				
0-38	74	143	22	296	68
41	Salyrgan i.v.				
64-94	70	1160	59	1370	60
97	✱6063 i.v.				
108-138	54	1133	375	1737	68
139	BAL i.m.				
154-203	63	113	362	564	146

Each datum represents mean of 3 consecutive periods.

TABLE 4. ELECTROLYTE EXCRETION IN CONTROL PERIODS AND AFTER SUCCESSIVE ADMINISTRATION OF ✱6063, SALYRGAN AND BAL (*Dog P*)

Time	C_{cr}	Excretion of			
		Cl^-	HCO_3^-	Na^+	K^+
min	ml/min	μEq/min.	μEq/min.	μEq/min.	μEq/min.
-65	Start infusion of creatinine, phosphate(pH 7.4), saline.				
0-40	117	20	50	603	152
47	✱6063 i.v.				
57-93	109	323	691	1724	298
96	Salyrgan i.v.				
118-151	112	843	692	2455	148
156	BAL i.m.				
176-208	121	896	648	2475	285

Each datum represents mean of 3 consecutive periods.

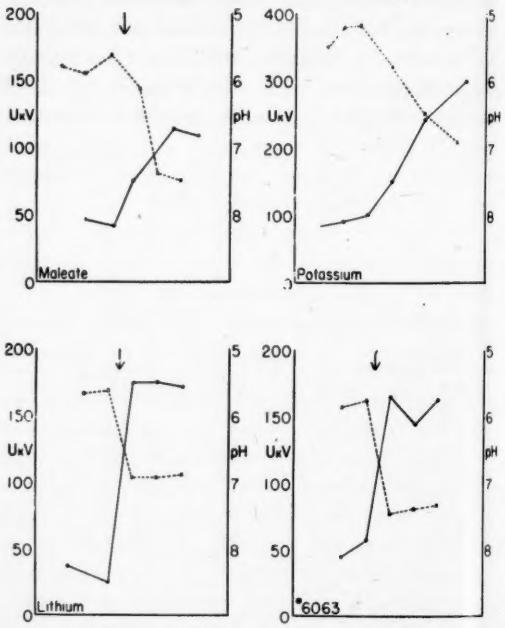


Fig. 2. URINE pH (*dotted line*) and potassium excretion (*solid lines*) before and after administration of indicated substances (see text).

perimental abnormalities which have been noted, but all are compatible with this explanation.

The very large amounts of bicarbonate excreted after ✱6063 is administered have been noted without comment. A brief consideration of this observation seems desirable since it appears to require some modification of the generally accepted view concerning the reabsorption of bicarbonate by the tubules. It is generally believed that reabsorption in the proximal tubule is not accompanied by a change in the pH of the fluid (12, 26); the development of an acid pH is considered a unique function of the distal segment. Because of the differences between their effects on the pH of the fluid, these have been considered to be fundamentally different proccesses—that in the proximal segment a bicarbonate-reabsorbing mechanism, that in the distal tubule, the carbonic anhydrase-dependent sodium for hydrogen ion exchange mechanism which has been under examination. A consideration of the relationship between plasma bicarbonate and the excretion of bicarbonate and titratable acid has led to the conclusion that, in acidosis, little if any bicarbonate reaches the distal hydrogen ion secreting mechanism and that at the peak of bicarbonate reabsorption the distal mechanism is responsible for only about 20 per cent of the bicarbonate reabsorbed (27). The type of data shown in table 5 indicates that revision of at least part of this picture is necessary. This table presents data from an experiment in a severely acidotic dog given a moderate dose of ✱6063. The plasma bicarbonate was then progressively elevated by infusion of sodium bicarbonate and the amounts filtered and excreted were measured. Despite the severity of the acidosis, appreciable amounts of bicarbonate were excreted throughout the course of the experiment.

TABLE 5. EFFECT OF ✱6063 ON EXCRETION OF BICARBONATE AT VARYING LEVELS OF PLASMA BICARBONATE (*Dog M, 19.3 kg.*)

Time min	Urine Flow ml./min.	Creatinine Clearance ml./min.	Plasma HCO_3^- μEq./min.	Filtered HCO_3^- μEq./min.	Excreted HCO_3^- μEq./min.	Excr./Filt. HCO_3^-
0	Priming: Creatinine 2.0 gm.; 6063 20 mg./Kg.					
1	Infusion: Creatinine 13 mg./min.; 6063 35 mg./Kg./hr. in 0.85% NaCl at 5.2 ml./min.					
87	Replace 0.85% NaCl with 2.0% $NaHCO_3$					
87-109	2.45	49	9.7	481	86	.18
109-127	6.70	52	12.2	649	214	.33
127-147	7.35	57	12.9	764	302	.40
147-167	7.21	56	15.0	869	382	.44
167-188	6.34	56	17.6	1020	431	.42
188-210	6.45	56	18.4	1081	496	.46
210-232	6.97	56	(19.5)	(1142)	585	(.51)
232-252	7.10	60	20.5	1280	596	.47

Figures in parentheses based on an interpolated value of plasma bicarbonate.

Although the fraction of the filtered which was excreted was somewhat lower in the first two than in subsequent periods, it is not certain to what extent this is due to the low plasma bicarbonate or to a somewhat lower rate of glomerular filtration. In any case, the rate of bicarbonate excretion is far too great to be attributed to even complete abolition of the distal hydrogen ion secretory mechanism. Either bicarbonate reabsorption is inhibited in the proximal as well as the distal tubule or else bicarbonate must be secreted by the tubules. Despite a diligent attempt, no evidence has been adduced to indicate that bicarbonate is secreted. It therefore seems highly probable that bicarbonate reabsorption in the proximal as well as the distal portion of the tubule is impaired when carbonic anhydrase inhibitors are administered. The question arises, then, as to whether these are really distinct mechanisms. The only reason for such a belief would appear to be that in one case pH change is believed to be small, the other large. This, however, could well be attributed to a difference in anion transport in the two regions. If anion reabsorption, particularly chloride reabsorption, were to keep pace with sodium transport, no net acidification would occur, whereas if no mechanism for transport of anion were present, exchange of sodium for hydrogen ions would result in acidification (see below). It is reasonable to proceed on the assumption that bicarbonate reabsorption throughout the renal tubule is a manifestation of hydrogen ion secretion in exchange for sodium. Until specific evidence of a second and different mechanism is adduced, it seems logical to concentrate study on a single process which is sufficient to account for the observed phenomena.

To extend the same line of reasoning somewhat

further, one may consider the possible implications of this mechanism as it concerns the reabsorption of sodium throughout the renal tubule. The essential trend of the material which has been under consideration has been one of breaking down distinctions: the distinction between the mechanisms for secreting hydrogen ions and potassium ions, and the distinction between bicarbonate reabsorption in the proximal tubule and hydrogen ion secretion in the distal tubule. It seems in order to reexamine the distinction between that sodium reabsorption which has been under consideration and the remainder of sodium reabsorption.

The chemical characteristics of the sodium ion are such as to make it difficult to conceive of a reaction into which this ion might enter that does not involve utilization of its electrical charge. Barring demonstration of some entirely new chemical property of sodium, one must conclude that it can be actively transported only by the displacement of some other ion of like charge (ion exchange). It would appear most reasonable, then, that acidification, potassium secretion, and bicarbonate reabsorption represent manifestations of a more general ion exchange process for the reabsorption of sodium.[2] Some additional support for such a view may be derived from the fact that one group of substances known to affect sodium transport, specifically the adrenal steroids, has been found to increase not only acidification of the urine and the excretion of potassium, but also that part of sodium reabsorption which is accompanied by the reabsorption of chloride.

REFERENCES

1. McCance, R. A. and E. M. Widdowson. *Lancet* 2: 247, 1937.
2. Keith, N. M., A. E. Osterberg and H. E. King. *Trans. Assoc. Am. Physicians* 55: 219, 1940.
3. Berliner, R. W. and T. J. Kennedy, Jr. *Proc. Soc. Exper. Biol. & Med.* 67: 542, 1948.
4. Mudge, G. H., J. Foulks and A. Gilman. *Proc. Soc. Exper. Biol. & Med.* 67: 545, 1948.
5. Mudge, G. H., J. Foulks and A. Gilman. *Am. J. Physiol.* 161: 159, 1950.
6. Pitts, R. F. and R. S. Alexander. *Am. J. Physiol.* 144: 239, 1945.
7. Foulks, J., P. Brazeau, E. S. Koelle and A. Gilman. *Am. J. Physiol.* 168: 77, 1952.
8. Van Slyke, D. D., A. Hiller and B. F. Miller. *Am. J. Physiol.* 113: 611, 1935.
9. Berliner, R. W., T. J. Kennedy, Jr. and J. G. Hilton. *Am. J. Physiol.* 160: 325, 1950.
10. Berliner, R. W., T. J. Kennedy, Jr. and J. G. Hilton. *Am. J. Physiol.* 162: 348, 1950.
11. Pitts, R. F., W. D. Lotspeich, W. A. Schiess and J. L. Ayer *J. Clin. Investigation* 27: 48, 1948.
12. Pitts, R. F. *Federation Proc.* 7: 418, 1948.
13. Darrow, D. C., R. Schwartz, J. F. Iannucci and F. Coville. *J. Clin. Investigation* 27: 198, 1948.
14. Darrow, D. C. *New England J. Med.* 242: 978 and 1014, 1950.
15. Loeb, R. F., D. W. Atchley, D. W. Richards, E. M. Benedict and M. E. Driscoll. *J. Clin. Investigation* 11: 621, 1932.
16. Winkler, A. W. and P. K. Smith. *Am. J. Physiol.* 138: 94, 1942.
17. Kennedy, T. J., Jr., J. H. Winkley and M. Dunning. *Am. J. Med.* 6: 790, 1949.
18. Broch, O. J. *Scand. J. Clin. & Lab. Investigation* 2: 113, 1950.
19. Schwartz, W. B. *New England J. Med.* 240: 173, 1949.
20. Miller, W. H., A. M. Dessert and R. O. Roblin, Jr. *J. Am. Chem. Soc.* 72: 4893, 1950.
21. Berliner, R. W., T. J. Kennedy, Jr. and J. Orloff. *Am. J. Med.* 11: 274, 1951.
22. Kennedy, T. J., Jr. Unpublished observations.
23. Berliner, R. W., T. J. Kennedy, Jr. and J. G. Hilton. *Proc. Soc. Exper. Biol. & Med.* 75: 791, 1950.
24. Foulks, J., G. H. Mudge and A. Gilman. *Am. J. Physiol.* 168: 642, 1952.
25. Orloff, J. and T. J. Kennedy, Jr. *Federation Proc.* 11: 115, 1952.
26. Montgomery, H. and Pierce, J. A. *Am. J. Physiol.* 118: 144, 1937.
27. Pitts, R. F. *Am. J. Med.* 9: 356, 1950.
28. Wilbrandt, W. *J. Cell. Comp. Physiol.* 11: 425, 1938.
29. Rapoport, S. and C. D. West. *Am. J. Physiol.* 162: 668, 1950.

[2] If sodium ions are transported by exchange for hydrogen ions, this process in itself would provide no electrical potential favoring the passive movement of anions. It must be assumed that most of the anion (chloride) is transported by a specific process, most probably an anion exchange (chloride for hydroxyl or bicarbonate) or that the electrical gradient necessary for passive movement of anion is provided by a process secondary to the sodium-hydrogen exchange (back diffusion of both these cation species but that of hydrogen at a higher rate). Few studies of the electrical potential across the renal tubule have been reported, but the potential difference found for the amphibian tubule was negligible (28). This, along with other evidences of specificity in anion transport (29) makes the exchange mechanism by far the more likely. To the extent that sodium and chloride reabsorption remain equal, no change in pH would be expected.

RENAL EXCRETION OF SODIUM AND WATER

HOMER W. SMITH

From the Department of Physiology, New York University, College of Medicine, New York City

ANALYSIS of the fate of sodium and water in the renal tubules may begin with the classic micropuncture studies of Richards and his coworkers on the amphibia. In 1937 Walker, Hudson, Findley and Richards (1) demonstrated in the frog and *Necturus* that the urine remains isosmotic with the plasma throughout the length of the proximal segment; since in the phlorizinized animal glucose undergoes some concentration in the proximal segment (2), it was inferred that sodium and water are absorbed there in a parallel or isosmotic manner. Subsequently, Walker, Bott, Oliver and MacDowell (3), from micropuncture studies in the guinea pig and rat and using creatinine as a standard of reference, concluded that by the end of the proximal segment some 80 per cent of the filtrate would be reabsorbed, but again isosmotically, as in the amphibia. Sodium with its attendent anions supplies some 90 per cent of the osmotic pressure of the glomerular filtrate, and such extensive reabsorption of water by the proximal segment must entail the proportional reabsorption of sodium, other substances contributing little to the osmotic picture. In 1948 Wesson and Anslow (4) adduced evidence from studies of mannitol diuresis in the dog that, in the over-all operation of the tubule, and unquestionably in the proximal segment, the reabsorption of sodium is an active process, analogous to the active reabsorption of glucose; whereas the reabsorption of water appears to be a passive process because a large fraction of the filtered water can be swept into the urine by the osmotic activity of the mannitol with only a slight increment in sodium excretion. They concluded that as sodium and other constituents are absorbed in the proximal segment, water is reabsorbed by passive diffusion and consequently the osmotic U/P ratio remains close to 1.0. Similar conclusions were reached by Mudge, Foulks and Gilman (5) from studies of urea diuresis in the dog (fig. 1).

Supplementing the view that the proximal reabsorption of water is a passive process, Wesson, Anslow and Smith (6) proposed that the thin segment of the loop of Henle serves as a region promoting passive osmotic equilibration before the urine enters the distal tubule where the final operations regulating sodium and water conservation are carried out.

Against the older view of Burgess, Harvey and Marshall (7), that the urine is concentrated to the hypertonic state in the thin segment, may be advanced the arguments, *1*) that its thin epithelium does not on cytological grounds suggest an operation of any such magnitude; *2*) casts, protein or foreign condensations of any kind are rarely observed here, but first appear in the ascending limb of the distal tubule and increase in density in the collecting ducts; *3*) in the human kidney, ⅞ of the nephrons have only a very short thin segment or none at all, and many of these do not descend into the medulla (8); and, *4*) one may ask rhetorically why should the urine be concentrated in the thin segment, at the expense of considerable osmotic work, before the final operations presumptively attributed to the distal tubule are effected? Therefore what is said below about distal function bears importantly upon our interpretation of function in the proximal segment and thin limb.

Wirz, Hargitay and Kuhn (9), on the basis of cryoscopic measurements in the frozen kidney, have recently proposed that the urine is actively concentrated, in the osmotic sense, in the extreme tips of the medullary loops, to be diluted again in the ascending limb of the loop of Henle. They conceive that the concentrated urine in the thin segments draws water across the interstitium of the medulla, the vasa recta and the collecting ducts, which must be supposed to be freely permeable to water, thus concentrating the urine in the collecting ducts. We may note that this hypothesis is opposed by all the arguments stated above, and also by the perhaps inadequate teleological argument that such a mechanism requires that all the blood flowing through the vasa recta must momentarily—i.e. for the period of its sojourn in the medulla—be concentrated to the same degree. Since probably no less than ⅕ of the renal blood flow traverses the medullary circulation, this consideration alone would make the operation a very extravagant one. We believe that the concept of isosmotic equilibration in the thin limb is in better agreement with the facts.

It is to the distal segment that we must look for the mechanism for making a dilute urine (fig. 2). This operation in the amphibia cannot consist of the tubular excretion of water, as Brodsky and Rapoport (10) suggest (for the mammal), because the urine as it passes down the distal segment becomes progressively more concentrated with respect to urea, phosphate, dyes and glucose in the phlorizinized animal (11); dilution must therefore be effected by the abstraction of isosmotically active material from the hitherto isosmotic urine, as originally accepted by Walker, Hudson, Findley and Richards (1). Direct evidence on distal function is not available in mammals, but indirect evidence indicates that here, too, dilution is effected by the abstraction of osmotically active material (T^d_{osm} in figure 2) from the isosmotic urine discharged from the proximal system.

It has long been recognized that maximal water diuresis in man represents about ⅛ of the water

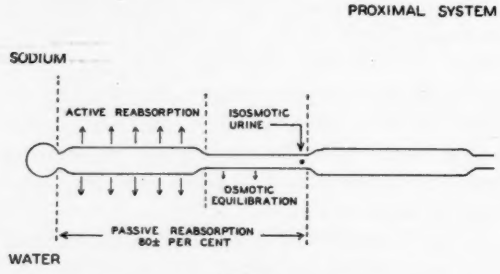


Fig. 1

of the glomerular filtrate. The only substance available in sufficient quantity to afford this quantity of osmotically free water by tubular reabsorption is sodium with its major attendant anions, chloride and bicarbonate. During water diuresis the urine may be so dilute in respect to sodium that we can fairly speak of it as almost sodium free, and the conclusion is unavoidable that this terminal step in the reabsorption of sodium is an active process.

By this active reabsorption of sodium, osmotically unobligated water is for the first time left within the nephron. The urinary clearance of osmotically free water, C_{H_2O}, is directly given by the difference between the total urine flow and the isosmotic clearance,

$$C_{H_2O} = V - \frac{U_{osm}}{P_{osm}} V$$

where U_{osm} and P_{osm} are determined by the cryoscopic method. The solute, expressed as osmols (T^d_{osm}), abstracted from the urine in order to create this free-water clearance, is given by the product of the free-water clearance times the osmotic pressure of the plasma. If one conceives that only sodium is involved in this operation, i.e., neglecting a possible minor contribution from potassium, the sodium chloride equivalent of this reabsorbed solute is given by dividing T_{osm} by f, the appropriate osmotic activity coefficient of sodium chloride (which we take to be 1.855).

Wesson and Anslow (12) have recently shown that, in *maximally hydrated* dogs, the free-water clearance, C_{H_2O}, is roughly constant, and the osmotic equivalent, T^d_{osm}, quite constant in any one animal during simple water diuresis and during extreme urea diuresis or natriuresis induced by mercurin plus cysteine. On their evidence we must conceive that so long as the load of sodium is adequate to effect saturation of the distal system, T^d_{osm} (or T^d_{Na}) is constant and

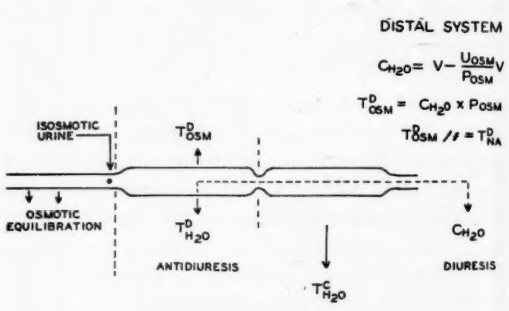


Fig. 2

maximal ($T^d_{m_{Na}}$). This constancy of the free-water clearance during variable osmotic diuresis argues against attribution of water diuresis to failure of water reabsorption in the proximal system.

This problem has been under further examination in man by Zak and Becker in our laboratory and they find that T^d_{osm} is fairly constant as between simple sustained water diuresis and variable mannitol diuresis superimposed on sustained water diuresis. Man, however, differs from the dog in that he presents greater difficulty in respect to attaining copious sodium excretion, and if sodium is not available to the distal tubule in minimal amounts the free-water clearance will be accordingly reduced; consequently at the present stage of these observations in man it cannot be said that distal electrolyte reabsorption is maximal, even though it is fairly constant in the same individual from test to test. Nor can it be

asserted, either for dog or man, that distal sodium reabsorption is removed from endocrine or humoral influence.

In the above view, during water diuresis the urine is composed of isosmotically obligated water (whether by sodium, urea, mannitol, or other solutes is immaterial) plus osmotically free water; the latter is made available by the absorption of electrolyte (presumably mostly sodium) distally. The dependence of gross water diuresis on electrolyte excretion, and the terminal reduction in gross diuresis in subjects who are continuously ingesting water, as reported by Rosenbaum, Nelson and Strauss (13), may be related to either reduced delivery of sodium to the distal tubule or to variations in urinary water isomotically obligated by sodium, rather than to changes in distal reabsorptive capacity ($T^d_{m_{Na}}$).

We have up to this point been referring to experiments in which sustained water diuresis was maintained by continued hydration of the body. In the absence of such hydration the osmotically free water liberated by the distal reabsorption of sodium is reabsorbed some place along the distal nephron, reducing the urine flow to oliguric levels. We tentatively suppose that it is reabsorbed *pari passu* with the sodium, possibly by passive diffusion through a facultatively permeable epithelium, leaving the urine again isosmotic with the plasma.

Under the action of the antidiuretic hormone, however, the urine passes above the isosmotic state to a variable degree of hypertonicity, and we now turn to the mechanism whereby an osmotically concentrated urine is elaborated.

The recent studies of Ladd (14) indicate that this hypertonicity is achieved by the reabsorption from the urine, regardless of the volume flow, of a constant quantity of water. In Ladd's experiments on prehydrated subjects, where sodium was the major osmotically active substance in the urine, this quantity amounts to about 2 cc. per 100 cc. of glomerular filtrate.

For convenience of discussion, this concentrating operation has been designated by Wesson and Anslow (12) as $T^c_{H_2O}$. There is as yet no direct evidence permitting us to decide whether $T^c_{H_2O}$ is functionally related to $T^d_{H_2O}$, or whether, indeed, the two processes even occur in the same part of the distal segment. One bit of evidence that they are functionally and anatomically separate is afforded by the clinical observation that concentrating power is frequently lost without loss of the capacity for water diuresis, and conversely during recovery from the ischemuric episode and other renal injuries, recovery of concentrating power is frequently long delayed. Moreover, to assign these two processes to a common locus requires that the tubule must be so constituted that, in the absence of the antidiuretic hormone (i.e., during water diuresis), it is virtually impermeable to water and yet capable of reabsorbing water against osmotic pressure during antidiuresis. While this versatility is conceivable, it seems improbable. The writer has elsewhere (15) presented the speculation that the concentrating operation, $T^c_{H_2O}$, may be located in the collecting ducts, and if so it may operate continuously and independently of the diuresis-antidiuresis mechanism.

Thus, the isosmotic urine entering the distal tubule is subject to two operations. The distal reabsorption of sodium leaves available a quantity of osmotically unobligated water, C_{H_2O},

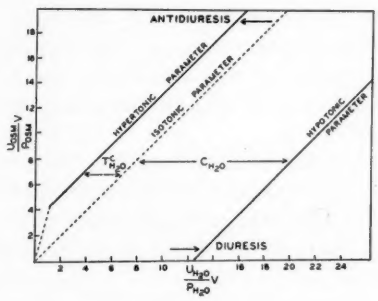


Fig. 3

which is reabsorbed during antidiuresis, but excreted during water diuresis (fig. 2). The magnitude of this free-water clearance as measured during sustained maximal water diuresis indicates that distal sodium reabsorption, T^d_{Na}, is characterized by a maximal rate and therefore the quantity of sodium delivered to the distal tubule is critical in determining both sodium excretion and the magnitude of water diuresis. The total urine flow, V, at any time represents the sum of the free-water clearance plus the isosmotic clearance of sodium, urea, mannitol and other constituents in the urine, thus generating the hypotonic parameter shown at the right of figure 3. The only way in which the free-water clearance could come to exceed the osmotic equivalent of distal sodium reabsorption would be in circumstances reducing water reabsorption proximally, so that free-water would be delivered by the proximal system to the distal tubule.

During antidiuresis, on the other hand, this

osmotically free water is reabsorbed, and during maximal antidiuresis a fixed quantity of water is reabsorbed from the now isotonic urine to generate the hypertonic parameter shown at the left of figure 3. Between these extremes of maximal water diuresis and maximal antidiuresis the urine may have any osmotic composition, reflecting the variable secretion of antidiuretic hormone and the variable excretion of osmotically obligated water.

The control of antidiuretic hormone secretion by the osmotic pressure of the plasma, classically elucidated by Verney and his co-workers, needs no discussion here. Attention may, however, be called to the recent evidence presented by Strauss, Davis, Rosenbaum and Rossmeisl (16) and by Leaf and Mamby (17) that critical changes in the volume of the extracellular fluid or, alternatively, in plasma volume or distribution (18), may mediate true water diuresis independently of changes in osmotic pressure. Similarly in Ladd's (19) prehydrated subjects, the infusion of isotonic saline solution leads to transient but frequently maximal water diuresis in the face of maintained or even increased osmotic pressure of the plasma.

Neither is it necessary to review the extensive literature showing that in the dog changes in extracellular fluid volume cause marked changes in filtration rate, a circumstance which most investigators believe plays a role in the conservation or excretion of sodium by disturbing glomerular-tubular balance. (Cort's recent paper (20) should, however, be mentioned.) In addition to the mechanism of glomerular-tubular balance, however, are changes in the reabsorptive activity of the tubules. That the tubular reabsorption of sodium is influenced by one or more adrenal hormones has been known since 1933 when Loeb and his associates demonstrated that the significant physiological disturbance in adrenal insufficiency is loss of sodium chloride through the kidneys. In 1935 Peters (21) suggested that the kidney retained osmotically active substances, chiefly sodium, when the extracellular fluid volume was depleted. Setting aside necessarily complicated experiments with orthostatic syncope (see Epstein *et al.* (22) for literature), evidence, however tentative, is accumulating that isosmotic expansion or contraction of the body fluids may lead to changes in the tubular reabsorption of sodium, as well as to changes in the tubular reabsorption of water, as noted above. These renal responses appear in an exaggerated form in the prehydrated subjects studied by Ladd (14, 23). Non-prehydrated subjects excrete isotonic saline solution sluggishly and variably (24) but if the subject is massively prehydrated 8–13 hours beforehand, the response to isotonic saline solution is now an abundant natriuresis coupled with the water diuresis mentioned above. Changes in the filtration rate are not significantly involved in this phenomenon. During the response to the infusion of saline solution, such prehydrated subjects excreted in peak periods an average of nearly 9 per cent of the filtered load of sodium, a figure comparable with the maximal effects of mercurial diuretics. Since the free-water clearance in such prehydrated subjects is not reduced below its average value of 12.5 per cent of the filtration rate, it is inferred that distal sodium reabsorption is not reduced and that this abundant natriuresis results entirely from reduction in proximal reabsorption. Expansion of the extracellular fluid volume appears to be the most obvious place to look for the stimulus evoking this copious natriuresis. The response of prehydrated subjects to isotonic saline solution is undoubtedly related to the changes in sodium excretion induced in non-prehydrated subjects by the excitation of volume receptors as described by Viar, Oliver, Eisenberg, Lombardo, Willis and Harrison (25) and by the expansion of the extracellular fluid, as described by Strauss, Davis, Rosenbaum and Rossmeisl (26).

In résumé, it may be emphasized that, supplementing the osmotic pressure of the plasma in controlling water diuresis, is the volume of the extracellular fluid, expansion or contraction of which may lead independently of plasma osmotic pressure to decreased or increased antidiuretic hormone secretion. Of paramount importance in the excretion of sodium is filtration rate and glomerular-tubular balance, and the critical barrier presented to the excretion of sodium by the distal tubule which serves to conserve sodium if the filtration rate is reduced or proximal tubular reabsorption increased. But of equal importance are changes in proximal reabsorption. The tentative interpretation of Ladd's data (14) indicates that in prehydrated subjects proximal reabsorption may be decreased as much as 10 per cent or more by isosmotic expansion of the extracellular fluid, thus greatly overloading the distal tubule despite constancy in the filtration rate. Whether this reduction in proximal reabsorption is mediated through the adrenal cortex is undetermined, nor is it known, as I have noted above, whether or

not distal sodium reabsorptive capacity, which at the moment has been measured only under very limited circumstances, is subject to humoral control.

I close with the hopeful note that the problems of sodium and water excretion appear to be approaching a new stage of integration. It is becoming clear that the renal operations involved in sodium and water excretion are integrated not only with the composition of the body fluids but also with their volume. The demonstration of this integration promises substantial progress in the near future for both the experimentalist and the clinician.

REFERENCES

1. Walker, A. M., C. L. Hudson, T. Findley and A. N. Richards. *Am. J. Physiol.* 118: 121, 1937.
2. Walker, A. M. and C. L. Hudson. *Am. J. Physiol.* 118: 130, 1937.
3. Walker, A. M., P. A. Bott, J. Oliver and M. C. MacDowell. *Am. J. Physiol.* 134: 580, 1941.
4. Wesson, L. G., Jr. and W. P. Anslow, Jr. *Am. J. Physiol.* 153: 465, 1948.
5. Mudge, G. H., J. Foulks and A. Gilman. *Am. J. Physiol.* 158: 218, 1949.
6. Wesson, L. G., Jr., W. P. Anslow, Jr. and H. W. Smith. *Bull. New York Acad. Med.* 24: 586, 1948.
7. Burgess, W. W., A. M. Harvey and E. K. Marshall, Jr. *J. Pharmacol. & Exper. Therap.* 49: 237, 1933.
8. Peter, K. *Untersuchungen über Bau und Entwickelung der Niere.* Jena: Gustav Fischer, I. 1909, II. 1927.
9. Wirz, H., B. Hargitay and W. Kuhn. *Helv. Physiol. Acta* 9: 196, 1951.
10. Brodsky, W. A. and S. Rapoport. *J. Clin. Investigation* 30: 282, 1951.
11. Richards, A. N. *Harvey Lecture* 93 : 1934–35.
12. Wesson, L. G., Jr. and W. P. Anslow, Jr. In preparation.
13. Rosenbaum, J. D., W. P. Nelson III and M. B. Strauss. *J. Clin. Investigation* 29: 841, 1950.
14. Ladd, M. *J. Applied Physiol.* 4: 602, 1952.
15. Smith, H. W. *The Kidney: Structure and Function in Health and Disease.* New York: Oxford Univ. Press, 1951, p. 321, 347.
16. Strauss, M. B., R. K. Davis, J. D. Rosenbaum and E. C. Rossmeisl. *J. Clin. Investigation* 30: 862, 1951.
17. Leaf, A. and A. R. Mamby. *J. Clin. Investigation* 31: 60, 1952.
18. Welt, L. G. and J. Orloff. *J. Clin. Investigation* 30: 751, 1951.
19. Ladd, M. *J. Applied Physiol.* 3: 379, 1951.
20. Cort, J. H. *J. Physiol.* 116: 307, 1952.
21. Peters, J. P. *Body Water. The Exchange of Fluids in Man.* London: Ballière, Tindall and Cox, 1935.
22. Epstein, F. H., A. V. N. Goodyer, F. D. Lawrason and A. S. Relman. *J. Clin. Investigation* 30: 63, 1951.
23. Ladd, M. *J. Applied Physiol.* 3: 603, 1951.
24. Crawford, B. and H. Ludemann, *J. Clin. Investigation* 30: 1456, 1951.
25. Viar, W. N., B. B. Oliver, S. Eisenberg, T. A. Lombardo, K. Willis and T. R. Harrison. *Circulation* 3: 105, 1951.
26. Strauss, M. B., R. K. Davis, J. D. Rosenbaum and E. C. Rossmeisl. *J. Clin. Investigation* 31: 80, 1952.

GASTRIC SECRETION OF ELECTROLYTES

FRANKLIN HOLLANDER

From the Gastroenterology Research Laboratory, The Mount Sinai Hospital, New York City

PURE gastric secretion—in contrast to stomach contents, which contain saliva, pancreatic secretion, bile, etc.—is itself a mixture of several secretions, derived from three or more distinct types of cells constituting the gastric gland proper: *1*) Parietal cells, source of the HCl; *2*) zymogen cells, source of pepsin and possibly other enzymes; *3*) columnar cells, source of mucus; and *4*) neck chief cells, source of the hypothetical mucoid secretion.[1] For purposes of analysis, this composite of gastric secretions may be subdivided into 2 components: An *acid component*, which consists only of the parietal secretion; and an *alkaline component*, which comprehends all the others together with interstitial fluid, desquamated cells, and gastric secretions of unknown origin such as the intrinsic factor.[2] Mixed gastric secretion contains all the electrolytes contained in other body fluids, including even bicarbonate on occasion, but their concentrations vary widely and with no apparent regularity. Consequently, my object is to describe the highly variable composition of gastric juice in terms of *a*) the chemical anatomy of the 2 components, and *b*) their ionic interactions following secretion.

Chemical characterization of the individual secretions following direct collection from their parent cells by micro-techniques is not yet feasible, and so it has been necessary for investigators to employ an indirect approach to this part of the problem. This entails study of variations in concentrations of the several electrolytes with collection time following stimulation, volume-rate of secretion, nature of the stimulus, etc. Interrelations among these variables are then determined by graphic and statistical techniques with evaluation of the parameters of the regression equations, and physiological interpretation of the intercepts, slopes, asymptotes, etc. Most of the data available for this purpose have been derived from experiments on the dog, but occasional observations on cat and man suggest that generalizations in the present context which have been derived from the former species are probably applicable to the latter as well, with no more than minor differences.

In 1932, I proposed a hypothesis to explain the variations in composition of gastric juice in terms of certain postulates concerning the chemical anatomy of the 2 components (2). This hypothesis may be formulated as follows:

1) The parietal secretion is a singular physiological entity, the chemical composition of which is fixed regardless of the stimulus. It contains HCl at a relatively constant concentration of 165 mN. It contains no combined acid, because protein, phosphate, etc. are absent. It contains chloride at a concentration equal to that of the acidity; i.e., solely as HCl and not at all as neutral salts. Consequently, it contains none of the common cations. Its water content (i.e., its total concentration) is such that its osmotic pressure is slightly greater than that of blood, but probably equal to that of the parietal cytoplasm.

2) The alkaline component is not a physiological entity, but is a mixture of 3 or more different secretions and other body fluids in variable proportions. It contains all the inorganic electrolytes of extracellular fluid, at concentrations which were tentatively postulated as being similar to those of interstitial fluid, but which are now recognized as being capable of considerable variation around these values. Thus, the total cation concentration is about 165 mN as is the total anion concentration; the concentration of Cl is never less than 100 mN but may be considerably greater, and the sum of HCO_3-ion and other negative ions varies accordingly. The alkaline component also contains a miscellany of organic substances, including pepsinogen and the gastric mucins.

3) All the variations in electrolyte concentrations of pure gastric juice result *primarily* from dilution and neutralization of the parietal secretion by the alkaline component on admixture of these 2 fluids in varying proportions, with at-

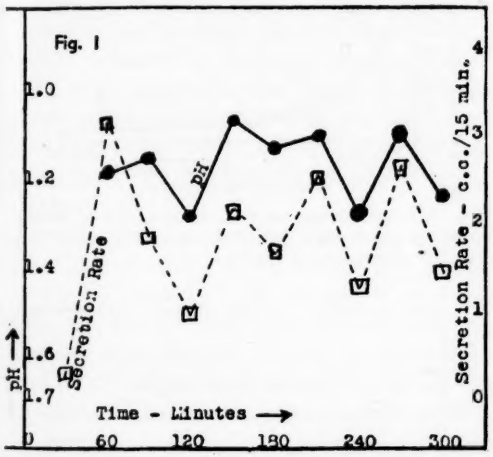
[1] Concerning the hypothetical nature of this secretion, see (1).

[2] The term 'non-parietal component' is used by some workers as synonymous with 'alkaline component.' The latter designation may be considered an exaggeration, since the pH of this mixture never exceeds 9. Nevertheless, this term is preferred to the former, not only because it suggests a pH greater than 7.0 but the fluid which it describes is potentially capable of playing an acid-neutralizing role in the stomach; also because 'non-parietal' may suggest a mixture of secreted fluids without some content of transudate and disrupted cells.

endant loss of CO_2; and *secondarily* from changes in osmotic pressure within the secreting cells themselves.

The evidence which I accumulated prior to 1938 gave very strong support to this 2-component hypothesis, and since then there has been considerable confirmatory evidence from other laboratories as well. However, there have also been observations in serious conflict with it, presented by investigators of unquestionable reliability. Hence, it is my purpose now to reexamine the major evidence bearing directly on this hypothesis, and to submit some new observations of our own[3] in refutation of those which have been advanced against it.

However, when the experimental procedure is modified so as to wash retained mucus out of the pouch before applying the stimulus, to minimize the proportion of mucus which is mixed with the parietal secretion during the experiment and to eliminate, if possible, all contamination by blood and exudate, such correlation between acidity and volume-rate frequently disappears. Thus, with a single maximal dose of histamine, the acidity curve rises more or less parallel with the volume until it reaches a maximum, but continues at this top level in spite of the subsequent falling off of the rate of secretion. In the experiment illustrated in figure 2, this constancy of acidity persisted for 105 minutes at a pH of 0.90 ± 0.01 (3). Such in-



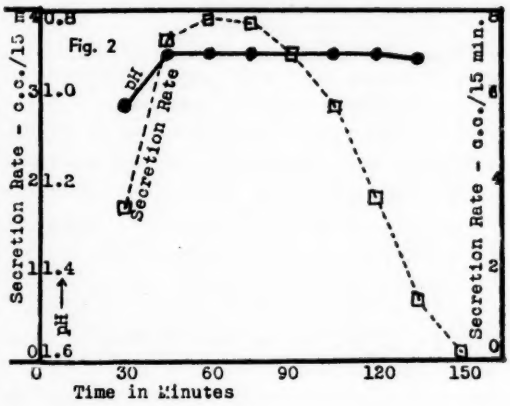


Fig. 1 (*left*). PARALLEL CURVES for acidity and rate of secretion, Pavlov pouch dog (*exper. B-5*, histamine).
Fig. 2 (*right*). NON-PARALLEL acidity and rate of secretion curves (constant pH), Pavlov pouch dog (*exper. B-27*, histamine).

RELATION BETWEEN ACIDITY AND VOLUME-RATE OF SECRETION

When pure gastric secretion is collected from canine pouches under ordinary experimental conditions, specimens being collected continuously for successive 15- or 20-minute periods following stimulation, the acidity is usually found to vary synchronously with the volume of the specimen, i.e., with the volume-rate of secretion. This parallelism is illustrated by the graph in figure 1 for a Pavlov pouch experiment which involved a single injection of histamine (3), but a similar relation has been reported for food, pilocarpine, and other stimuli by numerous investigators since the time of Pavlov. Such results imply a correlation of undefined mathematical character between the two dependent variables, acidity and rate.

[3] These new data derive chiefly from the work of my associates, Drs. Henry D. Janowitz and Henry Colcher, and I acknowledge my indebtedness to both of them.

dependence of the 2 variables was recently observed by us in a different type of experiment, one in which the stimulus was a small dose of histamine diphosphate (0.025 mg. base) given subcutaneously every 10 minutes to Heidenhain pouch dogs for 4–6 hours. As a rule, the volume-rate of secretion in such experiments is relatively constant; to offset this, a single injection of atropine was given at about the middle of the experiment, thus forcing the rate-curve down in spite of the continuous administration of histamine. Again, acidity and rate curves rise in parallel, after which the acidity stays relatively constant for at least 2½ hours while the rate falls off to a very low level (fig. 3).

Experiments like these agree well with the hypothesis that the parietal secretion has a constant acidity of 160–170 mN, and that the admixture of mucus results in deviations from this

maximum. This [HCl],[4] even in the absence of other electrolytes, has an osmotic pressure greater than that of canine and human blood under ordinary conditions. Such hyperosmotic concentration is probably a reflection of a similar characteristic of the parietal cytoplasm which results from the chemical changes entailed in the formation of the acid. Furthermore, such experiments indicate that when variations in acidity do occur they may be correlated with the volume-rate of secretion, but this relation occurs very irregularly, and may even be absent in suitably designated experiments. The notion that the parietal secretion has a constant acidity has not been questioned particularly, but the possible lack of correlation between the rate of formation of mixed juice and its acidity has been challenged most seriously. In

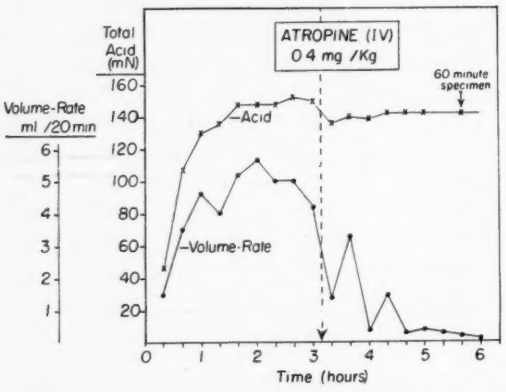


Fig. 3. GASTRIC secretory response to continuous histamine injection before and after atropine, Heidenhain pouch dog (*exper. HJ-38*). Stimulus: 0.025 mg. histamine base/10 min.

1934, Liu, Yuan and Lim (4) reported that there is a *unique* mathematical relation between the rate of secretion and the acidity of pouch juice following histamine stimulation—also, between rate and [Cl]. This relation is defined by a hyperbolic equation; its asymptote is determined by the constant maximum acidity of pure parietal secretion, and its rate-intercept affords a measure of the amount of alkaline component. Such hyperbolic relations were confirmed for pouch dogs by Gray, Bucher and Harman (5) and by Gudiksen (6) in acute experiments with cats. In fact, the hyperbolic regression technique has given promise of becoming a valuable tool for studying concentration-rate relations. A typical set of curves, illustrating the results of Gray *et al.*, are presented in figure 4 (7). All of these studies allow for gastric juice of high constant acidity, but at high rates of secretion *only*—corresponding to near-asymptotic values; at lower rates the acidity must invariably be expected to be low.

Our own finding, that high constant-acidities can be observed at low rates of secretion under suitable conditions, must stand as evidence in refutation of this last implication. Furthermore,

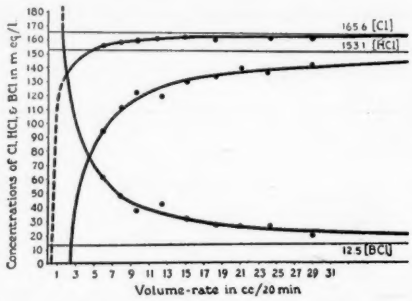


Fig. 4. RELATIONSHIPS between concentration of chloride fractions of gastric juice and volume-rate of secretion (after GRAY J. S. *Gastroenterol.* 1: 392, 1943).

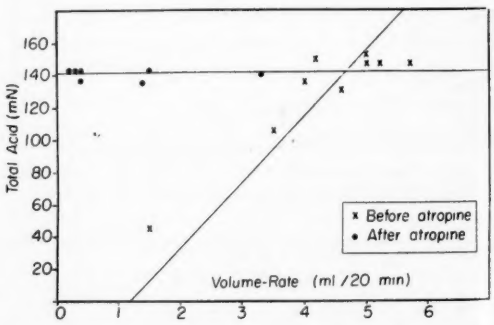


Fig. 5. GASTRIC secretory response to continuous histamine injection before and after atropine, Heidenhain pouch dog (*exper. HJ-38*). Stimulus: 0.025 mg. histamine base/10 min.; 0.4 mg. atropine/kg. body wt.

this particular application of the hyperbolic regression procedure is itself of dubious validity, because the observed regression lines and hence their equations do not always have the unique significance ascribed to them. This was demonstrated in crude fashion in 1931 (8) with some of our early data, but it can best be seen by replotting the acidity values of figure 3 directly against the corresponding rates. The resulting graph (fig. 5) demonstrates directly that the relation between these 2 variables is, to say the least, a 2-valued function; the horizontal portion

[4] The reader's attention is called to the use hereafter of a pair of brackets to designate 'concentration' of particular chemical individuals so enclosed.

corresponds to specimens which are nearly pure parietal secretion with a fixed minimal proportion of alkaline component, the remainder to specimens which are mixtures of both components in varying porportions. If the ratio of the 2 components changes greatly in the course of an experiment which is of sufficient duration to yield a large number of samples or points for this correlation analysis, these points may sometimes be scattered throughout the area included between the two straight lines shown in the illustration. This expectation follows also as a consequence of the interpretation placed by Liu *et al.* upon the rate-intercept, as reflecting 'the volume of the base mixture,' i.e., the proportion of mixed mucus, pepsin, etc. In other words, the data obtained in most such experiments may not be homogeneous in respect to the proportion of the 2 components, and the values for the parameters deduced from such data are liable to be spurious. Consequently, the physiological interpretation of these parameters is of questionable validity.

Additional experimental support for our finding that the acidity is not invariably a function of the volume-rate of secretion can be derived from Gudiksen's own observations with pilocarpine and vagal stimulation. In contrast with histamine, both of these stimuli failed to yield a well-defined correlation between acidity and rate, an observation which is particularly significant in view of the fact that the latter stimuli evoke high outputs of pepsin and mucus, whereas the former does not. Also, the asymptotes for the acidity (i.e., acid chloride) and total chloride regression lines were found, both by Liu *et al.* and by Gray *et al.*, to differ by a statistically significant amount, but our own experimental evidence based on direct measurement of sodium and potassium (see below) fails to confirm the validity of this difference, which corresponds to the neutral chloride concentration, [BCl].

COMBINED ACIDITY

It is generally recognized that, in the absence of contamination, the combined acidity of gastric juice (i.e., the titer between pH 3.5 and 7.0, or the difference between free and total acidities) is low. Our determinations on a series of pooled samples of histamine-induced secretion indicate a negative correlation of this variable with the total acidity itself, the value approaching zero as the acidity approaches 165 mN (9). These findings are consistent with the interpretation that protein is absent from the parietal secretion and hence is derived wholly from the alkaline component—an inference which is confirmed by direct evidence on the content of total organic solids in these pooled specimens and their responses to protein tests. The percentage of organic solids was found to be correlated negatively with the acidity, and the biuret and Hopkins-Cole tests varied progressively from ++ to 0 as the observed acidity rose to a maximum of 158 mN. This combination of evidence establishes beyond any reasonable doubt that the combined acidity of gastric juice is derived entirely from its alkaline component.

THE ANIONS

At pH's less than 3.5, HCO_3-ion must obviously be absent from gastric secretion, but it has been found in specimens which are chiefly mucous or mucoid in character and contain no free HCl. For these reasons, it is generally accepted that this anion is derived solely from the alkaline component of the gastric secretion.

Inorganic phosphate has been reported to occur in very small or negligible concentrations at high acidities. In our experience, several pooled specimens of pure pouch juice with [HCl] equal to 155 mN or greater, failed to yield any evidence of its presence (9) so that the absence of phosphate ion from parietal secretion is a safe inference. Data regarding its concentration in the several fluids which constitute the alkaline component may prove of value in elucidation of the latter.

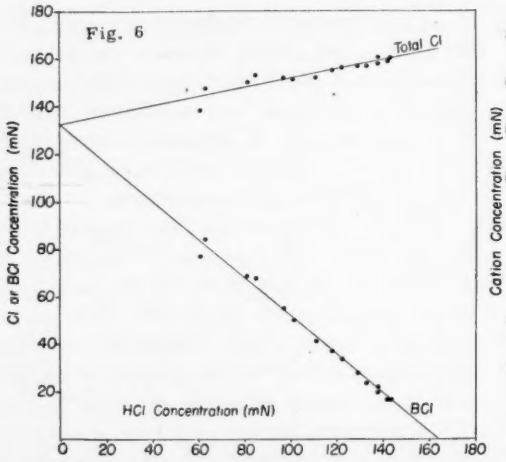
Of the several possible anions, therefore, only chloride demands much attention at present. Contrary to some early reports, the concentration of this element in mixed gastric secretion varies markedly, but it is invariably greater than the total acidity. Our hypothesis calls for [Cl] in pure parietal secretion equal to circa 165 mN; in the non-acid component it is capable of considerable variation between this value and 100 mN. When the latter component is predominantly interstitial fluid, transudate or a secretion of similar electrolyte pattern, its [Cl] may also be fairly constant near 100 mN; under these circumstances, [Cl] of the mixed gastric secretion may be expected to vary between this latter level and an upper limit of circa 165 mN, the regression with changing acidity being linear and of small positive slope (2). This type of correlation should characterize the picture in the upper part of the acidity range,

where the proportion of the alkaline component and hence the influence of its variable [Cl] is small. At lower acidities, and especially in the 'acid-deficit' region of high pH, where the proportion is relatively large and mucus seems to predominate, the [Cl] may no longer manifest this simple mathematical relation to the acidity.

Our own observations regarding the correlation between these 2 variables are entirely in accord with this expectation. Cell-free mucus of pH 7.4 or more (stimulated with acetylcholine), has yielded chloride concentrations of 113–151 mN (mean = 116 mN) which are wholly uncorrelated with the acidity (10). For fluid secretion of moderate to high acidity, obtained with histamine, however, there exists a well-defined linear correlation between its concentrations of total chloride and total (or even free) acidity, with a regression coefficient of small positive value (11). This kind of linear correlation has been confirmed by other investigators (5, 6, 12) severally for dog, cat and man; and for histamine, food and vagal stimulation. Earlier reports from other laboratories regarding parallel changes in the time curves for acidity and chloride for such experiments lend added weight to these findings, even though they lack statistical analyses. This linear relation between the 2 variables is illustrated in figure 6, constructed by us from the raw data of 2 dog experiments which MacLean *et al.* (13) published as demonstrating *constancy* of the total chloride concentration. How often this rectilinear function between [Cl] and acidity appears in the lower half of the acid range has not yet been investigated systematically, but its frequent occurrence seems most improbable in the light of the extensive variations in [Cl] which we have already reported from mucous secretion alone.

NEUTRAL CHLORIDE

The concentration of neutral chloride, represented by [BCl], is defined as the difference between [Cl] and [total acid]; and it has generally been employed as a measure of total base, i.e., of the total cation exclusive of H-ion. In that part of the acidity range where [Cl] is directly and linearly correlated with [HCl] our hypothesis requires that [BCl] must also vary with this acidity, but in inverse manner. The observations of many workers (e.g., 5, 11, 13, 14) conform closely to this expected relation. In the illustrative graph constructed from the data of MacLean *et al.*, the linear correlation coefficient is −0.994 and the intercept on the acidity axis 164 mN. For the individual sets of data reported by the other investigators, the latter statistic varies between 157 and 174 mN, with an average in the neighborhood of 165 mN. Since the correlation is unequivocally linear, and the intercept at the upper end of the acidity range corresponds to a [BCl] of 0 mN, it follows that the latter marks the lower limit to the variations in total base. From this, in turn, it may be inferred that this limit can be attained in actuality only in parietal secretion free of all non-parietal fluids, and could such *pure* acid secretion be collected for study,



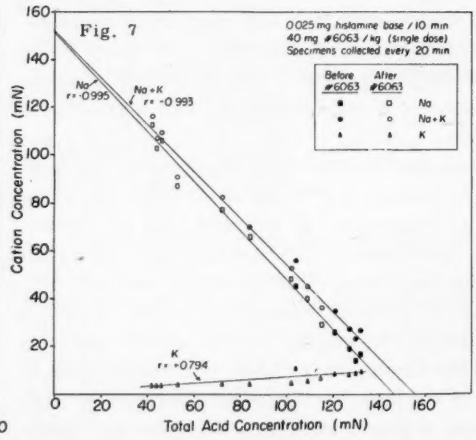


Fig. 6 (*left*). RELATIONS of total and neutral chlorides to acidity (from data of 2 Pavlov pouch experiments by MACLEAN, GRIFFITHS AND WILLIAMS).

Fig. 7 (*right*). RELATION of sodium and potassium concentrations to acidity in gastric juice, Heidenhain pouch dog (*exper. HJ-117*).

it would be found to be characterized by [HCl] = circa 165 mN and [BCl] = 0 mN. This inference is in accord with the hypothesis; its only weakness arises from the notorious unreliability of extrapolation methods generally. However, in our own experiments, values for [BCl] as low as 3–5 mN have been observed repeatedly, and Toby (15) cites numerous values equivalent to 6 mN in her tables and one observation corresponding to 1 mN in the text. Hence, the present argument rests on extrapolation for only a very short fraction of the entire acidity range, and since the extrapolation is rectilinear, the inference therefrom is acceptable without further qualification.

Nevertheless, serious evidence in refutation of this inference has been advanced by Liu *et al.* and subsequently by Gray, Bucher and Harmon; also by Fisher and Hunt, using the data of Ihre (16) on man for the purpose. In these three studies, the lower limit for [BCl] was reported as 12, 12.5 and circa 10 mN respectively, estimated as the difference between the asymptotic values for total Cl and BCl in experiments like that of Gray illustrated in figure 4. Since these estimates rest on the use of the hyperbolic regression technique, and the application of this statistical procedure to the present data has been shown above to be of questionable validity, this evidence regarding the magnitude of the neutral chloride content of parietal secretion can be given but little credence. Complete rebuttal—in addition to the observations of BCl values less than 10 mN already cited—is afforded by analysis of the situation in terms of the individual cations, especially potassium, which follows.

SODIUM

Analyses of the variations in [Na] in the course of experiments with various modes of stimulation have never revealed any indication that this variable is functionally related to the time following application of the stimulus, but it is so related—and inversely—to the acidity of the specimen. The relation between these 2 variables is characterized by a straight line, as in the illustrative experiment by Colcher and Hollander (17) shown in figure 7. Practically identical graphs have been reported by other investigators (cf. fig. 8) (6, 7, 12, 18–22). The similarity of this line to the one for BCl versus acidity in figure 6 is striking, and it should be noted that Na values as low as 3–5 mN have been found for histamine juice by Gray and Bucher (23) and Gudiksen, as well as ourselves. Consequently, it is now generally accepted that Na is absent from pure parietal secretion and that its concentration in the alkaline component is no less than that in blood serum. Our finding that [Na] in canine cell-free mucus is in the range 140–166 mN is in excellent agreement with this (10).

CALCIUM AND MAGNESIUM

Calcium also varies inversely with the acidity, as originally reported by Grant (24), and its exclusive derivation from the alkaline component has been confirmed by Gray and Bucher for dogs; Gudiksen for cats; and Kirsner and Bryant (25), as well as Rudd (26) for man. In none of these

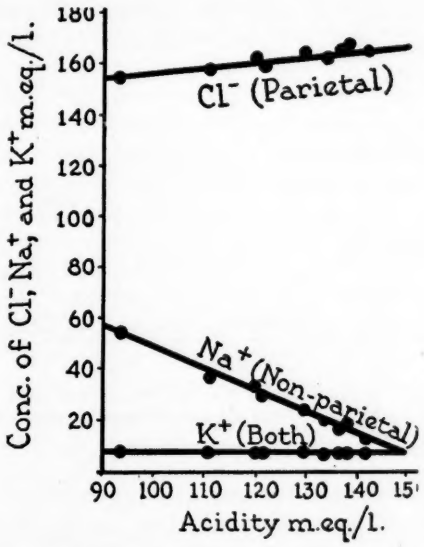


Fig. 8. Relationships between concentration of chloride, sodium and potassium (after Gray, J. S. *Gastroenterol.* 1: 393, 1943).

studies has a value for [Ca] greater than 2–3 mN been reported, except that of Rudd who found concentrations as high as 4.3 mN. As yet we have no data concerning the Ca content of acellular mucus, but our results for eugenol-stimulated mucus (which has a high cell content) and Grant's for viscous pyloric secretion are essentially the same as those for interstitial fluid—about 5 mN.

For magnesium, the reports are exceedingly scant, but those there are indicate that the concentration of this element in mixed gastric secretion is very low. Gudiksen has found this to be true also for secretion obtained by stimulation with pilocarpine or the vagus. Data on correlation of [Mg] with acidity or any other variable are

lacking, but it is the general opinion that this element is absent from the parietal secretion (6, 7, 27, 28).

POTASSIUM

Since Na, Ca, and Mg are all absent from parietal secretion, one might expect the same to be true for K—especially if the evidence in favor of no BCl in parietal fluid be accepted also. However, in the first systematic study of this element in gastric juice to be published, Gray and Bucher (23) reported that [K] followed a pattern entirely different from that of the other cations, being statistically constant at 7.4 mN, i.e., independent of acidity as well as volume-rate of secretion (fig. 8). From this the authors concluded that K occurs in both components of gastric secretion for subjects without disease of the digestive tract (12, 19, 27, 28, 32) and 0.5–65.0 mN with various types of gastrointestinal disorder (20, 27, 28, 30, 32, 33, 34).

In an effort to resolve the conflict between the several reports of constant [K] on one hand and this mass of variable and high data on the other, we have undertaken a systematic study of the variations in Na and K concentrations in gastric secretion. In Heidenhain pouch dogs, following a single subcutaneous injection of histamine (17), we found that [K] begins to rise immediately after administration of the stimulus, reaches a maximum within the first 45 minutes, and then declines steadily to a level which it maintains for the remainder of the experiment. This behavior, illustrated in figure 9, is fairly general, although

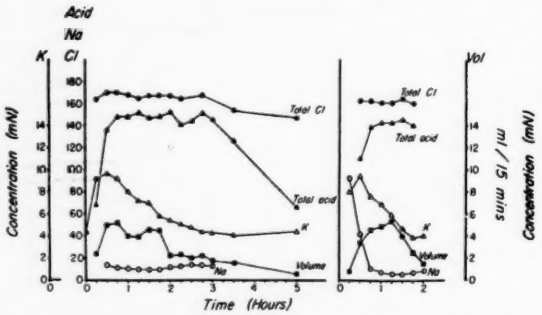


Fig. 9. CONCENTRATIONS of electrolytes in gastric pouch juice following single injection of 0.15 mg. histamine base/kg. body wt. (*expers. HC-M1, HC-E3*).

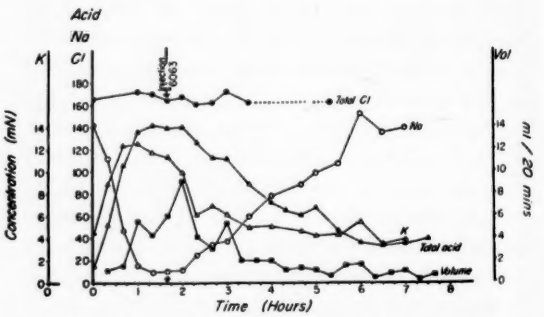


Fig. 10. CONCENTRATIONS of electrolytes in gastric pouch juice with repeated injection of histamine, 0.025 mg. base/10 min.; compound #6063, i.v., 30 mg/kg. body wt. (*exper. HJ-114*).

and that 7.4 mN represents its concentration in both of them. This inference derives support from the finding by Fisher and Hunt that [BCl] in parietal secretion is constant (at a level of 10 mN), and their suggestion that this electrolyte may consist solely of KCl. The data of Linde and Öbrink (29) *appear* to confirm this generalization also (at a concentration of 13 mN), though they believe that their evidence is not at all conclusive regarding such consistency (personal communication). Martin (30) and Gudiksen obtained no evidence for constancy of [K]; nevertheless their observations suggested to them also that the source of this element may be the parietal cell.

On the other hand, the literature contains many reports in conflict with Gray's. For dogs and cats, values well above 7.4 mN have been reported by several investigators with some values 2–3 times this concentration (6, 18, 20, 22, 27, 29, 31). For man, the variations are even more striking, falling in the range 3.6–35.5 mN some of the experiments show minor deviations which still require clarification. The lowest values attained fell in the range 0.4–5.0 mN and the peak values ranged as high as 11.3 mN. When 2 such experiments were performed in rapid succession, the second did not differ essentially from the first. The same pattern was manifest also by the K-curves in another series of experiments (35) in which histamine was injected in low dosages every 10 minutes for 6–8 hours, in order to maintain secretion at a relatively steady rate. After a suitable control period to establish stability of the volume-rate of secretion, the carbonic anhydrase inhibitor, compound #6063 (American Cyanamid Co.), was injected in order to reduce this rate during continuing administration of histamine (36). Again the K-curve reached a maximum within the first hour (fig. 10) and then fell off slowly to a more or less uniform value at a low level, this characteristic behavior being uninfluenced by the reduction in secretory rate

following administration of the inhibitor. The overall range for the [K] in 16 such experiments was 1.7–12.8 mN, not particularly different from that for the single histamine injection group.

These observations are diametrically opposed to Gray and Bucher's report of constant [K]. The spurious character of the latter now becomes evident from a reexamination of the procedure by which it was attained; all specimens collected during the first hour following initiation of the 10-minute histamine injections were discarded in order "to flush accumulated mucous secretion from the stomach." Furthermore, their study was conducted with 9 *pooled samples*, obtained by mixing 183 20-minute samples of gastric juice collected from 6 dogs according to the volume-rate of secretion. In this way, all the collection samples with high K-values were eliminated, and, by pooling the remainder in an effort to minimize random errors, small differences which were correlated with collection time were masked completely. If all our data were treated similarly —e.g., if the data for the 2nd and 3rd hours of each experiment were pooled and a mean (weighted by volume) calculated for this restricted set of observations—the result would not be significantly different from the values of 7.4 mN.

It is curious that the variations in [K] are uniformly correlated neither with acidity, volume-rate, nor [Na] (figs. 9 and 10). This becomes strikingly evident from scatter diagrams for these pairs of data. Such correlation may be encountered in an occasional experiment, but it is clearly fortuitous. Hence, there appears at this time to be little likelihood that K derives from the parietal secretion, especially since Dr. Colcher has found values as high as 42 mN in acid-free gastric aspirates from patients with pernicious anemia, nor does the evidence demonstrate its source to be the alkaline component. The only systematic variation in [K] is a time-trend which suggests some relation to the initiation of the secretory process, but not its continuation. The source of K in the pure gastric secretion is being investigated further.

CONCLUSION

The discussion herein presented leads to the *inescapable* conclusion that the *parietal secretion*, as it issues from its parent cell, is a solution of only hydrochloric acid which possesses an average concentration of 165 mN and is slightly hyperosmotic with respect to blood. None of the cations in mixed gastric secretion, other than H-ion, derive from this source. Chloride, bicarbonate, phosphate, sodium, calcium, and probably magnesium derive from the complex mixture which has been designated the *alkaline component* of the gastric secretion, but the source of the potassium is still unknown. Further exploration of the electrolyte pattern of gastric secretion calls for study of the variations in make-up of this alkaline component in the light of numerous factors which influence them.

Thus, the 2-component hypothesis which we have advanced to explain the variable and seemingly erratic character of the chemical anatomy of pure gastric secretion continues to be supported by all the major evidence currently available. Had the objection to this concept of the singular character of the parietal secretion been sustained by the evidence, it would have necessitated abandoning our theory concerning the mechanism of HCl formation by the parietal cell (37, 38). Since there is still every reason to believe that the parietal secretion is entirely free of cations other than H-ion, the concept that HCl-formation depends upon a membrane hydrolysis of neutral chlorides at the intracellular canalicular wall, with complete retention of all metal ions, is still tenable. However, need for amplification of the simple 2-component hypothesis is already evident, in order to account for wide variations in the concentrations of chloride and other anions in the alkaline component as well as the peculiar pattern of potassium concentration, and to allow for the possible occurrence of reabsorption and ionic interchange of sufficient magnitude to affect the electrolyte pattern of the mixed juice *after* ejection of its component secretions from their parent cells.

REFERENCES

1. Hollander, F. *Gastroenterol.* 20: 512, 1952.
2. Hollander, F. *J. Biol. Chem.* 97: 585, 1932.
3. Hollander, F. and G. R. Cowgill. *J. Biol. Chem.* 91: 151, 1931.
4. Liu, A. C., I. C. Yuan and R. K. S. Lim. *Chinese J. Physiol.* 8: 1, 1934.
5. Gray, J. S., G. R. Bucher and H. H. Harman. *Am. J. Physiol.* 132: 504, 1941.
6. Gudiksen, E. *Compt. Rend. Lab. Carlsberg (Serie Chimique)* 27: 145, 1950.
7. Gray, J. S. *Gastroenterol.* 1: 390, 1943.
8. Hollander, F. *Am. J. Physiol.* 98: 551, 1931.

9. Hollander, F. *J. Biol. Chem.* 104: 33, 1934.
10. Hollander, F. *J. Nat. Cancer Inst.* In press.
11. Hollander, F. *J. Biol. Chem.* 125: 161, 1938.
12. Martin, L. *Ann. Int. Med.* 6: 91, 1932.
13. MacLean, H., W. J. Griffiths and B. W. Williams. *J. Physiol.* 65: 77, 1928.
14. Fisher, R. B. and J. N. Hunt. *J. Physiol.* 111: 138, 1950..
15. Toby, C. G. *Quart. J. Exper. Physiol.* 26: 45, 1936.
16. Ihre, B. *Acta Med. Scand.* 97: Supp. 95, 1938.
17. Colcher, H. and F. Hollander. In preparation.
18. Gamble, J. L. and M. A. McIver. *J. Exper. Med.* 48: 837, 1928.
19. Schairer, E. *Klin. Wochenschr.* 8: 1113, 1929.
20. Austin, J. H. and G. D. Gammon. *J. Clin. Invest.* 10: 287, 1931.
21. Kozawa, S., K. Fukushima, M. Umeno, K. Kurihara, C. Takata and M. Horiuchi. *Jap. J. Med. Sci.* 3: 15, 1933.
22. Takata, C. *Jap. J. Med. Sci.* 3: 33, 1933.
23. Gray, J. S. and G. R. Bucher. *Am. J. Physiol.* 133: 542, 1941.
24. Grant, R. *Am. J. Physiol.* 132: 467, 1941.
25. Kirsner, J. B. and E. J. Bryant. *Am. J. Dig. Dis.* 6: 704, 1939.
26. Rudd, G. V. *Med. J. Australia* 1: 431, 455, 1934.
27. Bliss, T. L. *Ann. Int. Med.* 3: 838, 1930.
28. Saemundsson, J. *Acta Med. Scand.* 130: Supp. 208, 1948.
29. Linde, S. and K. J. Öbrink. *Acta Physiol.* 21: 54, 1950.
30. Martin, L. *Gastroenterol.* 15: 326, 1950; *Southern Med. J.* 43: 921, 1950.
31. Ingraham, R. C. and M. B. Visscher. *Proc. Soc. Exper. Biol. & Med.* 30: 464, 1933.
32. Dunham, L. J. and A. Brunschwig. *Cancer Res.* 6: 54, 1946.
33. Katsch, G., F. Baltzer and J. Brinck. *Arch. Verdauungskrankh.* 56: 1, 1934.
34. Myers, W. P. L. and J. W. Kirklin. *Proc. Staff Meetings Mayo Clinic* 27: 94, 1952.
35. Colcher, H., H. D. Janowitz and F. Hollander. *Federation Proc.* 11: 26, 1952.
36. Janowitz, H. D., H. Colcher and F. Hollander. *Federation Proc.* 11: 78, 1952.
37. Hollander, F. *Gastroenterol.* 1: 401, 1943.
38. Hollander, F. *Science* 110: 57, 1949.

SUBSTRATE AND OXYGEN CONSUMPTION DURING GASTRIC SECRETION[1]

HORACE W. DAVENPORT

From the Department of Physiology, University of Utah College of Medicine, Salt Lake City

ONE of the aims of the sciences of physiology and biochemistry is to describe completely the chemical reactions which result in the performance of a particular physiological function. This has not been accomplished in any instance, but in default of complete information it is useful to strike a balance between the quantities of substrate and oxygen used by a tissue and the amount of chemical or physical work done. In the case of chemical work done in gastric acid secretion it is desirable to calculate the molar efficiency given by the ratio of the number of hydrogen ions secreted to the number of molecules of substrate or oxygen used. Such a calculation provides a framework within which an acceptable theory of the nature of the secretory process must be erected.

In considering both substrates and oxygen used, the important point must be remembered that when the gastric mucosa secretes acid it uses large quantities of substrate and oxygen. The quantities used are so great that the gross efficiency of the mucosa is low, and for this reason almost any postulated mechanism of acid secretion is possible. Of course, it is unlikely that all the substrate and oxygen is used for secretion, and the problem is to partition them between the non-secretory and the secretory processes. This cannot be done experimentally with any degree of certainty, and the partition must be made on the basis of assumptions and on the internal properties of the data. This is a difficult task, and consequently the interpretation of the data presented here is subject to revision if a more certain partition can be made.

One of the 4 known substrates supporting acid secretion is acetoacetate. Using an *in vitro* preparation of mouse stomachs we have been able to measure the amount of acetoacetate used by the stomach and the amount of acid secreted (1). The results are shown in figure 1. In the graph at the left the rates of acid secretion with or without stimulation by the parasympathomimetic drug carbaminoylcholine are plotted against the concentrations of acetoacetate in the salt solution. The lower curve shows that when the stomach is not stimulated the rate of acid secretion is entirely independent of the substrate concentration, and the upper curve shows that when the stomach is stimulated the rate of secretion is a positive linear function of substrate concentration until a maximum is reached. In the middle graph the amounts of acetoacetate disappearing from the salt solution in which the stomachs are incubated in the presence and absence of stimulation are plotted against substrate concentrations. When the stomach is stimulated, the rate of disappearance of the substrate is a linear function of substrate concentration. When the stomach is not stimulated, the relation is curvilinear and at the higher substrate concentrations falls a little below the straight line obtained in the presence of the stimulating drug. What becomes of the acetoacetate is at present unknown. Very little of it is reduced to beta-hydroxybutyrate, and none goes to acetone. We have not yet been able to measure the other products of its metabolism, and consequently we have not been able to correct the changes occurring in the pH of the buffer for any effect which might be produced by the disappearance of the carboxyl group of the acetoacetate molecule. The graph on the right is a combination of the other two. The rates of acid secretion are plotted against the amounts of acetoacetate disappearing. The lower straight line, obtained in the absence of stimulation, shows that the acid secreted does not increase when more acetoacetate is used. The upper line shows that in the presence of the stimulating drug about 1.2 molecules of acid appear for each molecule of acetoacetate disappearing.

These data seem to mean that when the stomach is stimulated to secrete acid the metabolism of the substrate is qualitatively changed. Observe on the right hand graph that when 5 molecules of acetoacetate disappear the unstimulated stomach secretes 6 molecules of acid while

[1] Unpublished work cited in this paper was supported by a grant from the United States Public Health Service, National Institutes of Health. The assistance of Lawrence C. Weaver, Virginia J. Chavré and Lowell A. Woodbury is gratefully acknowledged.

the stimulated stomach secretes 14. The same amount of substrate is used, but the physiological results differ by a factor of more than 2. The efficiency of the secretory process cannot be calculated simply by comparing the acid secreted with the substrate used. Each intermediate step between disappearance of the substrate and the appearance of the acid must be identified, and unfortunately we are not able to do this.

The relation between oxygen consumption and acid secretion was determined using a sac of frog gastric mucosa prepared by the method of Davies (2, 3). The sac was filled with phosphate-buffered salt solution and was placed in the same solution

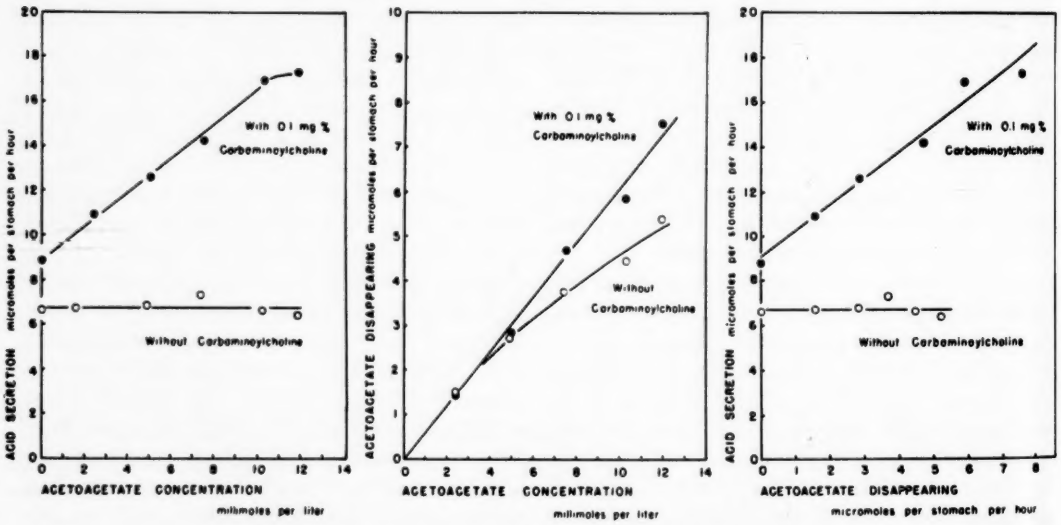


Fig. 1. ACETOACETATE UTILIZATION and acid secretion by mouse stomachs *in vitro*.

in a manometer flask. The oxygen consumption was measured over a period of 2 or 3 hours, and the secretion was measured by titration at the end of the period. Histamine in several concentrations was used to stimulate acid secretion or in other instances was omitted. We have always found that unstimulated mucosae secrete acid, sometimes at a high rate.

The results of 147 such observations at a pO_2 of 610 mm. Hg are shown at the left of figure 2. The rates of acid secretion are plotted against the rates of oxygen consumption. The correlation coefficient is +0.86, and there appears to be a reasonably good linear relation between the variables. The line drawn through the points is a regression line calculated by the method of Bartlett (4) to take into account the fact that both variables are subject to error. Its equation is given at the top of the graph, and it has a slope of 2.15.

It could be argued that the statistically determined relation between the variables describes the efficiency of the secretory process. This is shown on the right of the figure. The intercept of the line on the x-axis may represent the most probable value for the rate of oxygen consumption by the processes not responsible for the secretion of acid. The slope of the line may be the ratio between acid secretion and oxygen consumed or the efficiency. That this is true under the conditions of these experiments is strongly indicated by two-stage experiments. In these the acid secretion and oxygen consumption were measured in the first stage. Then the sac was washed out, refilled with salt solution, and in the second stage the acid secretion and oxygen consumption were again measured. Unfortunately, this procedure of doing the experiment twice on the same mucosa has a small effect upon the relation between acid secretion and oxygen consumption. This is demonstrated by the results shown in figure 3. At the left are observations made at a pO_2 of 610 mm. Hg and with various concentrations of histamine. The regression line plotted has the slope of 2.75. On the right are the results obtained on the same mucosae in the second stage in which the experimental conditions were the same as in the first stage. The slope of the regression line describing the data is 2.88. The two slopes and the intercepts of the lines on the x-axis are very slightly different

in the two cases but the differences have a very important bearing upon the interpretation of results obtained in two-stage experiments. It is not legitimate to compare the results obtained in the first stage with those obtained in the second stage without taking into account the shift of the population caused simply by doing the experiment twice.

We have performed two-stage experiments in which no histamine was used in the first stage and was used in the second stage. The purpose of these experiments was to see whether the changes in acid secretion and oxygen consumption induced by histamine stimulation agreed with the value calculated from the slopes of the regression lines shown earlier. If they did we would have strong reason to believe that the slopes actually describe the efficiency of the secretory process. We were forced to do the experiments in two stages, for it is impossible to assume that unstimulated mucosae do not secrete acid. In our hands unstimulated mucosae are never completely quiescent. If there were no effect produced simply by doing the experiment twice, then the results obtained with unstimulated mucosae should fall at the lower end of the regression line and the results obtained with stimulated mucosae should fall at the upper end. Furthermore, the ratios of the changes in acid secretion and oxygen consumption should have a mean value equal to the slope of the regression line. Because doing the experiment twice has an effect upon the relation between the variables the situation is more complex but no less convincing. This is shown in figure 4. At the left are plotted the results obtained on unstimulated mucosae. A reliable regression line cannot be calculated from these data, for they are too closely bunched. However, they should agree with the first stage results shown in figure 3, and the line obtained from this population is drawn through the points. The line obviously fits the data as it should, and the arithmetical mean of all the points falls very close to the line. The results obtained on the same mucosa in the presence of histamine are given in the middle of the figure. The line obtained in the second stage experiments shown in figure 3 is drawn through the points, and the mean of the data falls exactly on this line. At the right of the figure the two means obtained without and with histamine are plotted, and the two regression lines are drawn as dashed lines. The line connecting these two means has a slope of 5.23. If one were not aware of the shift of the population produced by doing the experiment twice one would conclude that this value of 5.23 is the true efficiency of the secretory process. However, it is obvious that this value is an entirely false representation of the efficiency. It might have any numerical value depending upon the position of the means on their respective regression lines. When the effect produced by doing the experiment twice is eliminated, it can be seen that the slope of the regression line does represent the true ratio of acid secreted to oxygen used.

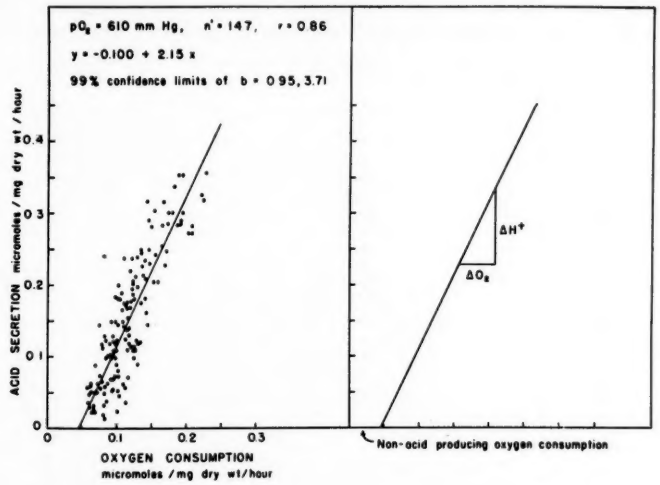

pO_2 = 610 mm Hg, n' = 147, r = 0.86
y = -0.100 + 2.15 x
99% confidence limits of b = 0.95, 3.71
ACID SECRETION micromoles / mg dry wt / hour
0
0.1
0.2
0.3
0.4
OXYGEN CONSUMPTION
micromoles / mg dry wt / hour
ΔH^+
ΔO_2
Non-acid producing oxygen consumption

Fig. 2. Oxygen consumption and acid secretion by frog gastric mucosae *in vitro*.

The slope of the line cannot be uncritically accepted as the maximum efficiency. The scatter of the data about the lines suggests that the efficiency may actually be variable. There are 3 possible reasons for the scatter of the data. *1)* There is undoubtedly observational error, but

confidence in our technique reenforced by actual measurement of the error leads us to reject the supposition that observational error is responsible for the whole of the scatter. *2*) There may be variations in the rate of oxygen consumption by the non-secretory processes, but we have no way of measuring this. *3*) Finally, the efficiency may actually be variable. Proof that it is variable is given in figure 5. On the left are those data obtained when the pO_2 was reduced to 125 mm. Hg in a one-stage experiment. It can be seen that under this condition the mucosae are capable of secreting acid at a high rate. The intercept on the x-axis is the same as that obtained with a higher tween the variables. The data do not fit the line describing the relation existing in the first part of the experiment, but they fall much farther to the left of it than can be accounted for by the effect of doing the experiment twice. They agree moderately well with the results shown on the left of the figure, for the two thin solid lines are those enclosing the data obtained in the one stage experiment under similar conditions.

These results prove that the efficiency is variable from one mucosa to another and that it is variable in the same mucosa from time to time. They also show that if the observational error were reduced to zero it would still be impossible

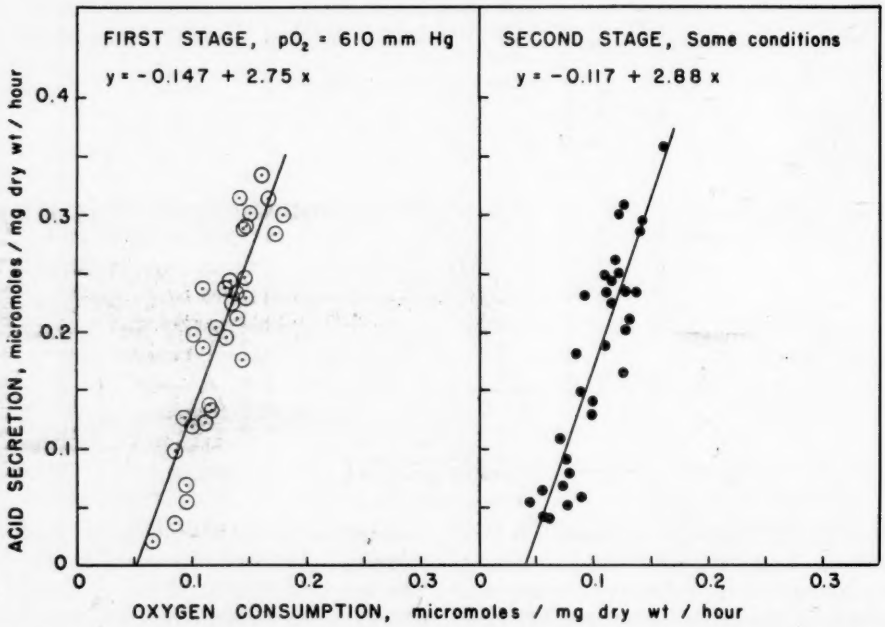


Fig. 3. Oxygen consumption and acid secretion by frog gastric mucosae *in vitro* in two stage experiments.

pO_2, but the slope of the regression line is significantly greater, being 3.57. Two thin lines are drawn to enclose all the points, and it is clear that the scatter of the points is narrow. At the right of the figure are plotted the results obtained in the second stage of a two-stage experiment. In the first stage the variables were measured with no histamine present and at a high pO_2. The results, which are not shown here, belong to the same population as that described in figure 1. The dashed line is the one obtained from this population. In the second stage of the experiment the mucosae were stimulated by histamine, and the pO_2 was reduced to 125 mm. Hg. The results show that these procedures changed the relation between the variables. The data do not fit the line describing the relation existing in the first part of the experiment, but they fall much farther to the left of it than can be accounted for by the effect of doing the experiment twice. They agree moderately well with the results shown on the left of the figure, for the two thin solid lines are those enclosing the data obtained in the one stage experiment under similar conditions.

These results prove that the efficiency is variable from one mucosa to another and that it is variable in the same mucosa from time to time. They also show that if the observational error were reduced to zero it would still be impossible to calculate the true efficiency of the secretory process solely from two pairs of observations on the same mucosa. This is shown graphically in figure 6. Suppose for the sake of simplicity it is assumed that the oxygen consumption of the non-secretory processes is constant at all times as shown by the intercept on the x-axis. Suppose, however, that the efficiency of the secretory process is not under absolute experimental control but that it may have either of two values whether the observer desires it or not. For example, the ratio of acid secretion to oxygen consumption may be either 2 or 4. These conditions are shown by the two slanting lines having a common intercept on the x-axis and slopes of 2

and 4. Then if acid secretion and oxygen consumption are measured first when the mucosa is unstimulated and second when it is stimulated, various apparent efficiencies can be obtained although in each instance the rate of acid secretion is increased by stimulation. One possibility is shown by *line A* in which the rate of acid secretion goes up as the result of stimulation but the rate of oxygen consumption goes down as it often does in fact. The apparent ratio in this case is −1.8. Other possibilities are shown by *lines B* and *C* whose apparent ratios are +22 and +0.8. It is clear that when the true ratio is variable the apparent ratio may have any value whatever. Since the ratio is equal to the tangent of the angle that the line joining the points makes with the x-axis, the ratio may vary from zero to plus infinity and from minus infinity to zero again. The situation is even more complex if the oxygen consumption of the non-secretory processes is variable and when there is observational error. We have performed many two-stage experiments and have calculated the apparent efficiencies. The results in effect lie in the whole range of possible values, and none of the values, be it 3 or 13, has any real meaning. The only two circumstances in which the efficiency can be measured by experiments of this kind are when the efficiency is known to be constant at any value or when it is known to be at its maximum. In either case other experimental evidence is necessary to establish the true efficiency.

We have not been able to design experiments which will establish the true efficiency independently of the slope of the regression line obtained for a large number of data. All that can be said is that in the experiments described and in many others we have never found the molar ratio of acid secreted to oxygen consumed to be significantly greater than 4. If this is actually the maximum ratio, the thermodynamic efficiency of the secretory process is only about 30 per cent. A ratio of 4 molecules of acid secreted to 1 molecule of oxygen used suggests that the underlying process is an oxidation-reduction reaction involving the transfer of one electron for each proton secreted. A number of hypothetical mechanisms involving such a process have been suggested by others (5), but they all suffer from the lack of experimental evidence to support them.

This analysis of the data rests upon the assumption that glycolysis does not contribute to acid secretion. If any hydrogen ions are secreted directly or indirectly as the result of glycolysis, they must be subtracted from the total acid measured before the ratio of hydrogen ions secreted as the result of oxidative processes to the molecules of oxygen consumed can be calculated.

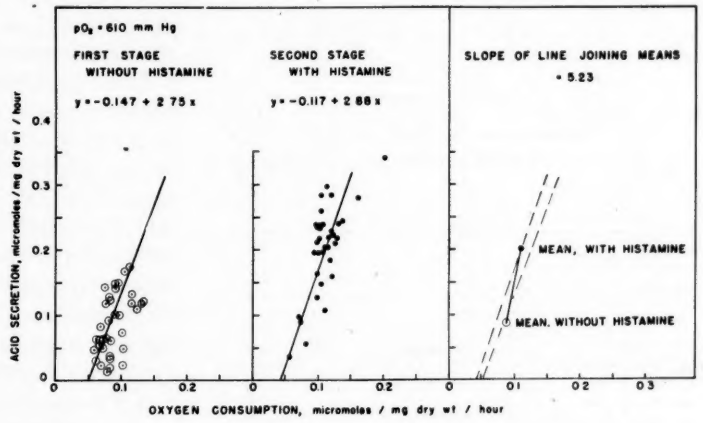

pO_2 = 610 mm Hg
FIRST STAGE
WITHOUT HISTAMINE
y = −0.147 + 2.75 x
SECOND STAGE
WITH HISTAMINE
y = −0.117 + 2.88 x
SLOPE OF LINE JOINING MEANS
= 5.23
MEAN, WITH HISTAMINE
MEAN, WITHOUT HISTAMINE
ACID SECRETION, micromoles / mg dry wt / hour
OXYGEN CONSUMPTION, micromoles / mg dry wt / hour

Fig. 4. OXYGEN CONSUMPTION and acid secretion by frog gastric mucosae *in vitro* in two-stage experiments. *First stage*, no histamine; *second stage*, histamine.

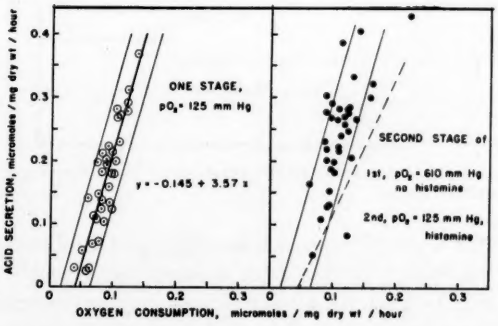

ONE STAGE,
pO_2 = 125 mm Hg
y = −0.145 + 3.57 x
SECOND STAGE of
1st, pO_2 = 610 mm Hg
no histamine
2nd, pO_2 = 125 mm Hg,
histamine
ACID SECRETION, micromoles / mg dry wt / hour
OXYGEN CONSUMPTION, micromoles / mg dry wt / hour

Fig. 5. OXYGEN CONSUMPTION and acid secretion by frog gastric mucosae *in vitro*.

Consequently, any contribution on the part of glycolysis would make the true ratio of acid secreted to oxygen consumed lower than the observed one. It has been argued by others (6) that energy derived from glycolysis cannot be used for acid secretion. The evidence is entirely that mucosae do not secrete under strictly anaerobic conditions and that when oxygen consumption is inhibited acid secretion stops. It remains possible

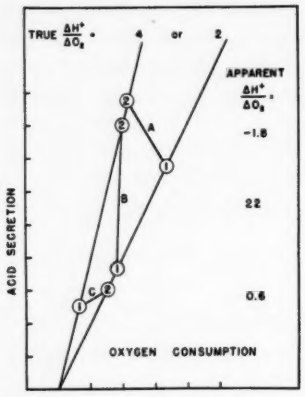


Fig. 6. INTERPRETATION of two-stage data when efficiency is variable.

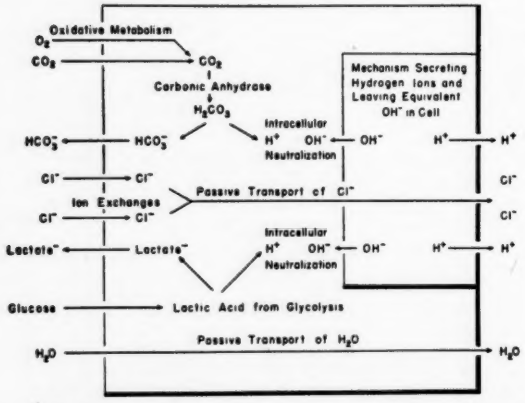


Fig. 7. INTRACELLULAR NEUTRALIZATION by carbon dioxide and lactic acid.

that when the mucosa is secreting acid aerobically the energy derived from concomitant glycolysis may be used. An aerobic process may be essential for secretion, but when that process is operating energy from glycolysis may be added to it. Since the rate of glycolysis increases as the partial pressure of oxygen falls, this would account for the observed apparent increase in the efficiency of oxygen utilization. We have made several experimental tests of this, but we have not been able to come to a definite conclusion. We have measured the rates of acid secretion and oxygen consumption by frog mucosae at two low partial pressures of oxygen when no glucose was present in the salt solution. Depriving the mucosae of glucose cuts their rate of glycolysis in half but does not abolish it. The results indicate that the gross oxidative efficiency of the mucosa may be lower when the rate of glycolysis is reduced. In addition, we have performed two-stage experiments in the first stage of which the mucosae were incubated in the absence of glycose and in the second stage of which they were given glucose. The results obtained could be explained by use of energy from the increased rate of glycolysis in the second part of the experiment. None of these groups of experiments is finished, but they suggest that it is possible that glycolysis does contribute energy for secretion. If that is true, the oxidative efficiency of the secretory process is still lower than that indicated by our data on oxygen consumption and acid secretion.

There is another way in which glycolysis is involved in acid secretion. It is well established by the work of Davies and his colleagues (7) that carbon dioxide is an intracellular neutralizing agent. This is shown at the top of figure 7. When acid is secreted an equivalent amount of alkali is left behind in the parietal cell. Carbon dioxide, derived from the metabolism of the cell and from the plasma, is hydrated to form carbonic acid, and carbonic anhydrase catalyses the reaction. The hydrogen ions subsequently formed by the ionization of carbonic acid neutralize the alkali, and the bicarbonate ions pass into the plasma in exchange for chloride ions. The consequences of this exchange are well understood (8). Our data show that lactic acid can fill the same role. We have found in more than 100 experiments that 97 per cent of the lactate produced by the mucosa passed into the solution on the serosal side of the mucosa and only 3 per cent into the lumen. This is true at all rates of secretion and all rates of glycolysis. It appears that the process shown at the bottom of the slide can also occur. Lactic acid is formed by glycolysis. Provided all other reactions associated with glycolysis such as the resynthesis of phosphocreatine are in a steady state, as they must be in indefatigable cells, the hydrogen ions derived from lactic acid are available for intracellular neutralization. The lactate ion then passes into the plasma in exchange for chloride. This process is particularly important in our *in vitro* preparations when they are operating in solutions nearly free of carbon dioxide and

in which the carbon dioxide produced by the cellular metabolism is absorbed by KOH. Consequently, intracellular neutralization by the carbonic acid system is reduced or eliminated. We conclude that carbonic acid and carbonic anhydrase are not indispensable but that they provide only the most convenient means within the parietal cell for accomplishing neutralization.

REFERENCES

1. Davenport, H. W. and V. J. Chavré. *Am. J. Physiol.* 166: 456, 1951.
2. Davies, R. E. *Biochem. J.* 42: 609, 1948.
3. Weaver, L. C. *Am. J. Physiol.* 167: 835, 1951.
4. Bartlett, M. S. *Biometrics* 5: 207, 1949.
5. Davies, R. E. *Biol. Rev.* 26: 87, 1951.
6. Crane, E. E. and R. E. Davies. *Biochem. J.* 49: 169, 1951.
7. Davies, R. E. and J. Edelman. *Biochem. J.* 50: 190, 1951.
8. Davenport, H. W. *Gastroenterol.* 1: 383, 1942.

ELECTROLYTE SECRETION IN THE PANCREAS

A. K. SOLOMON

From the Biophysical Laboratory, Harvard Medical School, Boston, Massachusetts

THE purpose of the present article is to review and summarize the literature concerning electrolyte secretion in the pancreas. Since the literature does not present a detailed hypothetical mechanism to account for pancreatic secretion, it seems desirable to assemble the available data in the form of such a hypothesis in the hope that it will stimulate further investigation. In this connection, calculations have been made of the energy required for pancreatic secretion in order to see whether sufficient metabolic fuel is available.

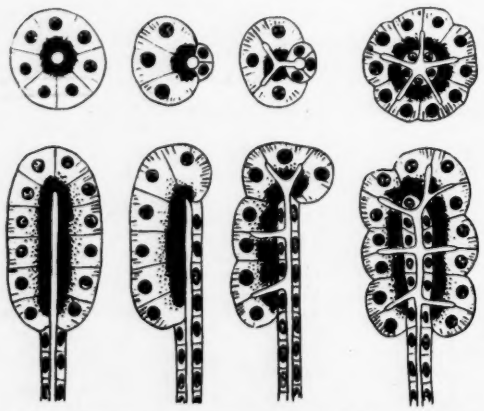


Fig. 1. DIAGRAMS of different relations of duct and gland cells in pancreatic acini; transverse sections above, longitudinal below (Zimmermann). (From BREMER, J. L. *Textbook of Histology Arranged Upon an Embryological Basis*, rewritten by H. L. WEATHERFORD, 6th ed. Philadelphia: Blakiston, 1944.)

Two types of pancreatic juices may be elicited by the action of hormones: one which appears following the administration of secretin is rich in electrolytes and poor in proteins; another which appears following the administration of pancreozymin is rich in proteins. We shall limit ourselves to the consideration of the electrolyte pattern of the secretin juice and will not be concerned with its protein content, if any.

The pancreas is composed primarily of three kinds of cells, acinar cells, duct cells and centro-acinar cells, in addition to the islets of Langerhans; a diagram showing the anatomical relation of the first two, which may be concerned with external secretion, is presented as figure 1. Histological studies have been interpreted to show that the protein secretion is a function of the acinar cells which comprise the greater part (75 per cent or more) of the pancreas. However, there does not seem to be any such evidence concerning the electrolyte secretion which may be formed by the acinar cells or the duct cells. Traditionally analyses of the chemical composition of the pancreas have been carried out on preparations from the whole gland, and these analyses have been of great use in elucidating the details of the mechanism of pancreatic secretion. In view of the great preponderance of acinar cells in the gland, this procedure amounts to a tacit assumption that they are responsible for the secretion of electrolytes. In the absence of analytical data on the electrolyte content of the separate types of cells, we shall also make the same assumption. Fortunately, in agreement with our assumptions, Lim, Ling, Liu, and Yuan (1) have suggested that the duct cells and centro-acinar cells do not contribute to the electrolyte secretion. However it should be explicitly understood that if the duct cells are eventually found to be the cells responsible, in whole or in part, for the production of the electrolyte secretion, all arguments based on the analysis of the whole gland will be invalid. One important problem in the elucidation of the mechanism of pancreatic secretion is the identification of the cells which actually secrete the juice, and the accurate analysis of their intracellular composition.

COMPOSITION OF PANCREATIC JUICE

The experimental animal which has been used most often in studies of the external secretion of the pancreas is the dog, and we shall consequently give figures that relate to the pancreatic secretion of the dog, except where otherwise noted.

It has been known at least since the studies of DeZilwa (2) in 1904 on freezing point depression in pancreatic juice and plasma that pancreatic

juice is isotonic with the blood. In 1933, Gilman and Cowgill (3) confirmed the observations of DeZilwa using a different method, and further found that induced alterations of the osmotic pressure of the plasma were followed very shortly by equal alterations in the osmotic pressure of the pancreatic juice.

The cationic composition of the juice in dogs has been measured by Johnston and Ball (4) and Ball (5). The concentrations of the principal cations are as follows: Na, 154 ± 7; K, 4.8 ± 0.9; Ca, 1.7 ± 0.3; Mg, 0.5 mmols/kg. H_2O. The anionic composition of the juice depends upon the rate of secretion in that the proportion of HCO_3 in the juice increases with increasing secretion rate as was pointed out by Ball (5) in 1930. This relationship was demonstrated quantitatively by Lim, Ling, Liu, and Yuan (1) in 1936, who showed that the sum of Cl and HCO_3 was constant, an observation which was confirmed by Komarov, Langstroth and McRae (6). The sum of the HCO_3 and Cl concentrations has been shown by Ball (5) and Johnston and Ball (4) to equal 154 ± 10 mmol/kg. H_2O, and the PO_4 to equal 0.8 ± 0.4 mmol/kg. H_2O.

Hart and Thomas (7) have studied the relationship between HCO_3 concentration and secretion rate extensively; we have plotted some of their data in figure 2. It will be seen that the HCO_3 concentration rises smoothly with the rate of secretion and approaches an asymptote at a concentration of 166 mEq/l. It should be pointed out that though investigators appear to agree that the juice is isotonic with the plasma, a wide variance exists between investigators with respect to the exact concentration of ions present in isotonic dog plasma, or pancreatic juice. For consistency we shall adopt the values of Ball and his collaborators in our calculations, although these values cannot, of course, apply to the data in figure 2.

Lim and his colleagues (1) suggest that the electrolyte secretion consists of 2 separate juices, one a neutral chloride at a concentration of 174 mEq/l., and another bicarbonate at a concentration of 174 mEq/l., and that the volume of the Cl concentration is always in a constant proportion to that of the HCO_3. It is not possible to distinguish between a mixture of 2 fluids of concentration 174 mEq/l., and a single fluid the sum of the concentration of whose anions is 174 mEq/l. However, the observation of Lim and his colleagues that the relative volume of these 2 fluids is always in the same proportion is not in agreement with the observations of Hart and Thomas. Lim *et al.* carried out their experiments at very low secretion rates, usually of the order of 1 or 2 ml/half hour, and plotted a time integral of the Cl content of the entire output against total volume secreted. This is a very insensitive index to changes in the Cl concentration arising from small differences in the rate of secretion, and may very well have masked the changes of the ratio of Cl concentration to HCO_3 concentration.

In addition to the 2 electrolyte fluids, Lim and his co-workers have indicated that there is a small amount of a third fluid, containing mucus, fer-

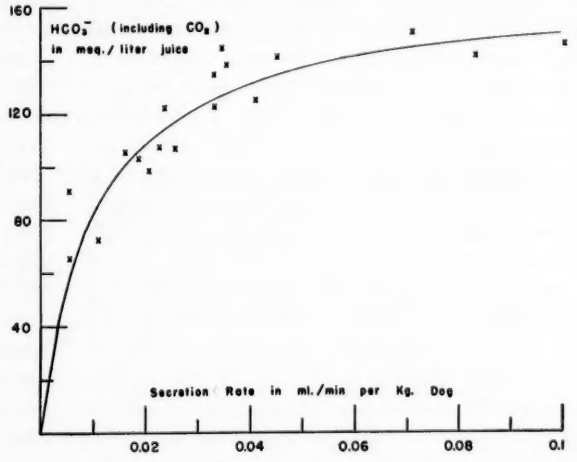


Fig. 2. HCO_3^- CONCENTRATION in pancreatic juice (data from HART AND THOMAS. *Gastroenterol.* 4: 409, 1945). *Dog I-44*, stimulus: glutamic acid; curve empirical, $(HCO)_3^- = \frac{166r}{r + 0.0106}$.

ments, and acid chloride. The presence of mucus and ferments makes it seem less likely that this fluid is concerned primarily with electrolyte secretion, since these components might well be expected to appear at the low rate of secretion used by Lim. Under these conditions it does not seem possible to assign a role to the acid chloride at the present time.

PRESSURE RELATIONS IN THE PANCREAS

Before considering the detailed mechanism of pancreatic secretion, it is necessary to examine the pressure relations within the gland. Although the normal pancreatic duct pressure appears to be very small, it has been shown by Berg and Zucker (8) that the pancreas can continue to secrete in the presence of back-pressures up to 30–35 cm. H_2O, which were the limiting pressures under their conditions. These pressures are very small compared with the osmotic pressure of the juice which is about 8000 cm. H_2O. Furthermore, Pavlov, as quoted by Babkin (9) carried out experiments in dogs in which, by an ingenious combination of venesection and vagal stimulation, he was able to obtain secretion pressures of the pancreas in excess of the venous pressure.

Additional support for the belief that the pressure relations are not important in the process of electrolyte secretion is gained from the observation of Chardon and Gross (10) that oxygen is apparently essential for the response to secretin of anaesthetized dogs. The volume of pancreatic juice secreted following a rapid injection of secretin fell 3-fold when the animal breathed 5 per cent O_2, and returned to normal when the O_2 concentration returned to normal. Further, Davies, Harper, and Mackay (11) who measured the Q_{O_2} of isolated pancreatic tissue from the cat found that secretin increased the Q_{O_2} by 18–36 per cent.

All this evidence may be taken as militating against the view that the pancreatic juice is formed by simple filtration, and we may be justified in considering that the pressure within the duct is unimportant in formulating a mechanism for pancretic secretion.

SOURCE OF IONS FOUND IN PANCREATIC JUICE

For many years a controversy has existed about the source of the ionic components of the pancreatic juice: the question of whether the ions come directly from the cells of the pancreas, or whether they are supplied directly by the ions in the blood. Since the pancreatic secretion in the dog comprises a considerable amount of fluid, about 250–350 ml/day according to Berg and Zucker (8), we require about 40 mmol each of Na and HCO_3 every day from a gland which has a wet weight of about 30 grams. Since such a supply is virtually impossible without some replenishment of the gland during the 24-hour period, it is clear the the question resolves itself into one of the time relationship. In 1930 Ball (5) showed that Na and K appeared promptly in the pancreatic juice following the injection of large amounts of these ions into the blood stream. We have taken some of his data and plotted them as figures 3 and 4, which show the rapid appearance of both Na and K in pancreatic juice following their intravenous injection.

Later, in 1940–41, Montgomery, Sheline, and Chaikoff (12) injected radioactive Na into the blood stream, and found that it appeared in the juice within 3 minutes following the injection. It is clear therefore, that any temporary loss of Na from the pancreas is made up relatively rapidly from the blood, and that the passage of an individual Na ion through the pancreatic cells must be complete in a matter of minutes. The experiments of Montgomery, Sheline, and Chaikoff agree well with the observations of Ball (5) in the case of Na. Consequently we would expect that the movement of K through the cells would be as rapid as that of Na, since the appearance of K in pancreatic juice following injection, shown in figure 3, is at least as prompt as that of Na shown in figure 4.

Further support for the view that the Na and K cannot all be drawn from the gland is found in the work of Oldfelt (13), some of whose results in the cat are shown in figure 5. It will be seen that after 3 hours of intensive pancreatic secretion, the Na content of the pancreas has fallen by only 14 per cent, which can account for only a small portion of the Na secreted in the juice during that time. Further, the pancreatic K has remained approximately constant. It should be pointed out that the cat is an animal with a serum osmotic pressure higher than that of the dog, and that therefore the sum of the cations in the cat pancreatic juice is greater than that of the dog.

The question of the origin of the pancreatic juice HCO_3 is complicated by the production of

CO_2 in the normal metabolism of the gland. In view of this situation, it has seemed more likely that the CO_2 in the pancreatic juice HCO_3 could originate in the gland, and as a consequence this supposition has been investigated at great length.

Still, Bennett, and Scott (14) measured the CO_2 content of blood before and after it had passed through the pancreas of an anaesthetized dog. They obtained rather variable results, since in 3 of 8 experiments the venous CO_2 fell below the arterial CO_2, and in the remaining 5 experiments it either equaled or exceeded the arterial CO_2. These results would seem to indicate that some of the CO_2 found in the juice may well originate from that formed metabolically.

The experiments of Still, Bennett, and Scott also provide further quantitative data from which it is possible to make a rough estimate of the amount of CO_2 produced by the pancreatic metabolism. They have shown that a continuously secreting gland consumes O_2 at a rate of 0.062 cc/gm. wet weight, and has an extra R.Q. of 1.05 during secretion. Combining these figures with the observation of Tucker and Ball (15) that an actively secreting 31 gm. pancreas of a dog can secrete 1.25 ml. of juice/minute, we can calculate that this O_2 consumption is accompanied by the production of 0.072 mols CO_2/l. of juice secreted. From figure 2 a concentration of 0.072 mols CO_2/l. of juice is equivalent to a secretion rate of 0.008 ml/min. kg. dog, or 0.12 ml/min. for a 15-kilo dog. From the paper of Still, Bennett, and Scott, it is possible to estimate that the secretion rate under their conditions was of the order of 0.2–0.3 ml/min. for a dog of about 14 kilos weight. Since these figures are not given directly by the authors, they must be accepted with considerable reserve and can serve as a general guide only. However, they suggest that the metabolic CO_2 can account at most for only about half of the CO_2 that appears in the juice.

Further confirmation for the view that much of the HCO_3 is drawn directly from the blood is afforded by the experiments of Ball, Tucker, Solomon, and Vennesland (16). They injected radioactive CO_2 intravenously in a dog and observed its prompt appearance in the pancreatic juice. The average ratio of total CO_2 in juice to serum which they observed was 4.75, and the average ratio of the radioactive carbon of juice to serum was 5.95. In view of the exigencies imposed by the short (20.5 min.) half-life of the C^{11} they used, the difference between the two ratios is probably not significant, and the results are seen to indicate that much of the CO_2 comes from the blood stream. Certainly the passage of the CO_2 from the blood to the serum is very rapid, since the ratio of C^{11} in juice to blood is as great in the first 10 minutes following injection as it is later in the experiment. However, in view of the inaccuracies of the technique, it can not be said

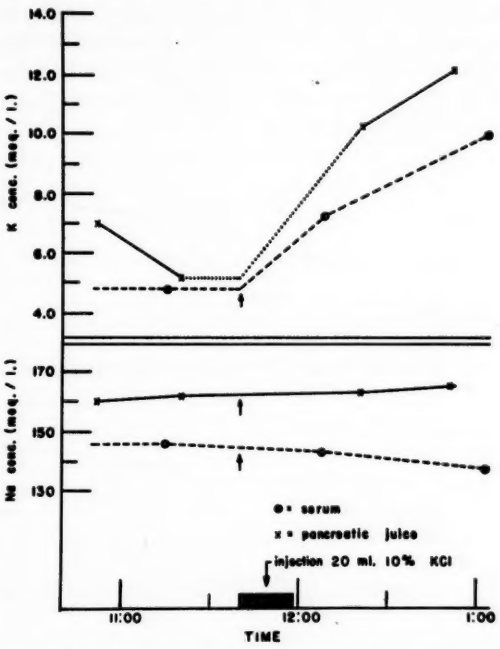


Fig. 3. CATION CONTENT of pancreatic juice (dog). (Data from BALL, E. G. *J. Biol. Chem.* 86: 449, 1930, table IV.) Experimental points are given by Ball. Dashed lines connecting serum points have been extrapolated by the present author on the assumption that serum levels are normally constant and change only following injection. The dotted interpolation between pancreatic juice points is based on a similar assumption. It has seemed instructive to connect these points by such interpolation, though it is realized that in the latter case it may well be in error.

that all of the CO_2 arises from the serum, and it is quite possible that it comes in part (perhaps 20% or less) from the metabolic CO_2 produced in the cells. Such a conclusion would seem the more likely since it does not seem inherently probable that all the metabolically produced CO_2 should go back to the blood stream in the presence of a rapid counterflow of CO_2 passing from the blood stream to the juice.

PERMEABILITY OF THE PANCREATIC CELL

Table 1, which shows the principal anions and cations in plasma, juice, and within the pancreas, illustrates the ionic environment of the organ. The intracellular data have been obtained from minced whole pancreas, and hence are primarily representative of the acinar cells. Since no data have been found in the literature which gives the intracellular concentrations of Na and K in the dog pancreas, we have carried out some measurements to provide these figures. These data are of a preliminary kind and should be taken as a general guide only. The serum data, except for the CO_2, are given by Eichelberger and Richter (17), and all other figures are taken from Ball and his colleagues (4, 5, 16).

in the plasma and juice. A similar situation exists in the cat pancreas. Oldfelt's determinations (13) of intracellular cations of secreting cat pancreas, as given in figure 5, are seen to be markedly different from the plasma concentrations of 5.3 mmol K/kg. H_2O, and 178.5 mmol Na/kg. H_2O as given by Fenn, Cobb, Manery, and Bloor (19).

With respect to the anions, Ball (15), and his collaborators have shown that Cl and HCO_3 appear in the cell at about 60 per cent of the plasma level, and hence have concluded that these ions move freely between the plasma and the interior of the pancreatic cells. However they suggest that the cell wall on the duct side is preferentially permeable to HCO_3, as compared with Cl, and consequently more HCO_3 appears in the

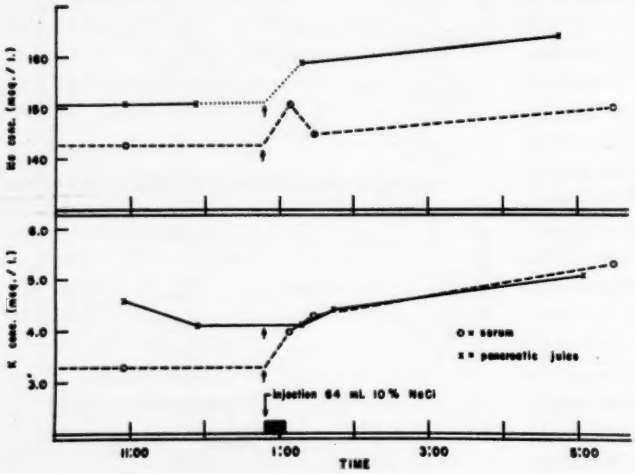


Fig. 4. CATION CONTENT of pancreatic juice (dog). (Data from BALL, E. G. *J. Biol. Chem* 86: 449, 1930, table I.) See caption for figure 3 for details concerning extrapolation and interpolation.

In view of the rapid appearance of Na and K in the pancreatic juice following injection, it had been previously assumed that the pancreatic cell was generally permeable to small cations. However, Ball (5) has shown that doubly charged cations such as Ca and Mg pass through much less freely than do the alkali cations. Furthermore, Visscher's work (18) on dyes does not seem to point to free cation permeability. Of the 85 dyes he tested, including equal numbers of cationic and anionic dyes, 35 were secreted by the pancreas, including 25 anionic dyes and 10 which were either amphoteric, or had structures similar to those of amphoteric dyes.

It is immediately apparent from table 1 that the acinar cells at least are not completely permeable to Na and K since the concentration of these ions is markedly different from that found

juice as the rate of secretion increases. Ball (5) has also shown that PO_4 does not enter the juice rapidly following injection into the blood stream.

Since some question exists concerning the preferential HCO_3 permeability, as will be shown below, and since the acinar cells do not appear to be cation permeable, it seems wise to reexamine the mechanism of pancreatic secretion in the light of these findings.

WORKING HYPOTHESIS FOR PANCREATIC SECRETION

Since one of the purposes of the present symposium is to promote discussion, it seems worthwhile to present a working hypothesis for the mechanism of pancreatic secretion, largely to serve as a basis for further examination and as a summary of facts known now. In view of the uncertainty about the nature of the cells which

are actually involved in the secretory process, it will be necessary to assume in the first instance that the acinar cells are actually responsible for the secretion, and to base our hypothesis upon this assumption. Then we will reexamine the hypothesis to see what modifications are necessary if the duct cells should be the ones actually involved.

Having ruled out hydrostatic pressure as an important factor in the mechanism of secretion, it is necessary first to determine how the water passes through the cell. The simplest and most logical assumption is that it is carried with the ions, and moves in a passive sense only. This would account for the isotonicity of plasma and juice, and for the rapidity with which changes in the tonicity of the blood are followed by similar changes in the pancreatic juice.

Since the Na and K are present in the cell in concentrations considerably different from those in blood and juice, we must assume that the cells are not permeable to these cations, but that some electrochemical energy is utilized in pumping the K up the concentration gradient from the plasma to the cell, and the Na up the concentration gradient from the cell to the pancreatic juice. It is clear that the system which is responsible for this cation transport is one which can distinguish between the two cations concerned because as Ball has shown (5), elevation of the concentration of either cation in the blood is followed by elevation of the concentration of the same cation in the pancreatic juice. Oldfelt's demonstration (13) of the striking fall in cellular Na in the pancreas of the actively secreting, secretin stimulated, cat would seem also to suggest that the Na concentration is intimately concerned with the secreting process. The cation balance in the cat is presumably maintained by an increase in H^+ concentration, since Hammarsten and Jorpes (20) have shown that the pH in the cat pancreas decreases from 6.74 for a resting pancreas to 6.41 for a secreting one.

The inverse relationship between the Cl and the HCO_3 concentration in the juice has already been discussed. The empirical equation from which the curve in figure 2 is drawn indicates that the secretion would consist entirely of HCO_3 at a concentration of 166 mEq/l. juice at the asymptote, and further that this asymptote would only be approached within 5 per cent at a secretion rate of 0.2 ml/min. kg. dog or 3 ml/min. for a 15-kilo dog. Since the maximum secreting rate is of the order of 1.5 ml/min., it seems likely that the Cl is always present in the juice, a feature which should be included in the working hypothesis.

Next let us turn to the suggestion of Ball, Tucker, Solomon, and Vennesland (16) that the cell wall on the duct side is less permeable to Cl than to HCO_3 and hence as the rate of secretion increases, the HCO_3 passes through the cell wall preferentially and consequently appears in increasing concentration in the juice. One index of

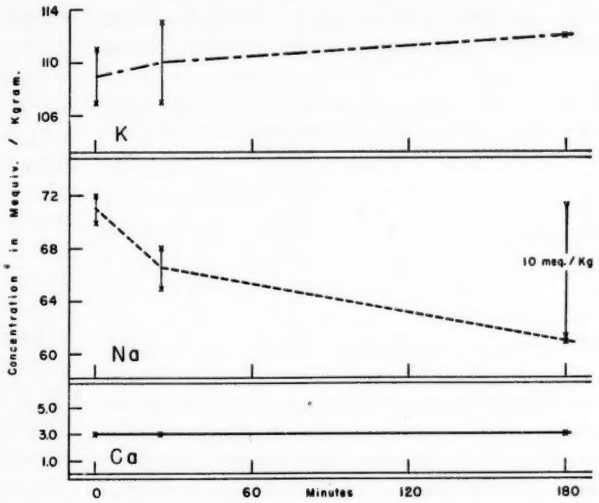


Fig. 5. CATION CONTENT of pancreatic tissue after intensive secretion (cat). (Data from OLDFELT, C. O. *J. Physiol.* 102: 362, 1943.) Experimental points are given by Oldfelt. Lines connecting them have been drawn through mean values.

the relative permeability to these ions is the limiting ionic conductance in water at 25°C., which bears an inverse relationship to the ease of ionic movement through water. Harned and Owen (21) give limiting ionic conductances (λ_0) in water at 25°C. for HCO_3^- of 44.5, for Cl^- of 76.34, and for OH^- of 197.6. Sutra (22) has calculated from similar data the following values of ionic radius: 1.20 Å for Cl^-, and 0.45 Å for OH^-. Using Sutra's formula and the data of Harned and Owen, we can calculate a value of 2.06 Å for the ionic radius of HCO_3^-, almost twice as large as that obtained for Cl^-. The ionic radius is not to be considered as having a real meaning, being rather the imaginary radius that an ion would have if it were spherical, had a uniform charge distribution and moved in an infinitely dilute solution; nonetheless, the ionic radius does serve as a qualitative index of ionic size. If we consider Fick's law of diffusion, and assume that the Sutherland-Einstein equation holds for the diffusion constant, we would expect the Cl to diffuse 1.7 times as fast as HCO_3 through the cell wall. However, in view of all the assumptions involved in this calculation, no reliance should be placed on the value of this ratio, which should be considered as a qualitative guide only.

TABLE 1. PRINCIPAL INORGANIC CONSTITUENTS OF SERUM, PANCREAS AND PANCREATIC JUICE IN DOG

	SERUM	PANCREAS	PANCREATIC JUICE
		mmol./kg. H_2O	
Cl	117.8 ± 4.8	70.9	154 ± 10
Total CO_2	28.3	18.2	
Na	153.1 ± 3.2	45 ± 4	154 ± 7
K	5.1 ± 0.3	105 ± 5	4.8 ± 0.9
H_2O	92.4%	77.1%	98%

The difficulties which this observation imposes on the suggestion of Ball and his colleagues disappear at once if we make use of three additional facts which are well known: the large lipoid solubility of CO_2, the presence of carbonic anhydrase in significant amounts within the pancreatic cell as reported by Davenport (23) and Roughton (24) and the small size of the OH radical. It is necessary further to make use of the observation of Davies and Roughton (25) that in the reaction scheme:

$$HCO_3^- + H_2O \rightleftarrows H_2CO_3 + OH^- \quad (1)$$

$$H_2CO_3 \rightleftarrows CO_2 + H_2O \quad (2)$$

Reaction 2 goes to the right on the acid side of pH 7.2 (actually $pa_H = 7.2$), and goes to the left on the basic side of this pH. The acid pH of the pancreatic cell observed by Hammarsten and Jorpes (20) would support the liberation of CO_2 within the cell, and the presence of carbonic anhydrase would accelerate this reaction. The CO_2 thus formed would diffuse very rapidly through the lipoid components of the cell wall. The OH radical, thanks to its small size, would also readily permeate the cell wall, and once outside would quickly rejoin the CO_2 to form HCO_3 in the alkaline medium of the juice.

Davies (31) has pointed out that the OH concentration in the alkaline juice is much greater than it is in the acid medium of the cell. Hence if the OH ion is to cross the wall it must either be actively transported against a concentration gradient, or more probably, be formed in a localized region of very high concentration just inside the cell wall, from which it can diffuse across the wall and into the juice. Although the alkaline medium of the juice favors the recombination of OH and CO_2, it may well be necessary that this recombination take place just without the cell wall within the sphere of action of the cellular carbonic anhydrase. It would further seem logical that *reactions 1* and *2* would take place just inside the wall on the duct side both in order to provide the local high concentration of OH, and to shorten the diffusion path for the CO_2. In this connection the 'cell wall' is taken to represent all the barriers between the secreting cell and the pancreatic juice, just as the 'cell wall' on the other side is taken to represent all the barriers between plasma and the secreting cell.

The experiments of Tucker and Ball (15), in which large amounts of carbonic anhydrase inhibitors injected into dogs failed to inhibit pancreatic secretion, may not be taken as evidence that the carbonic anhydrase system is not concerned with pancreatic secretion, since Tucker and Ball themselves did not believe that the carbonic anhydrase system was completely blocked. Further, the work of Davies (26) seems to indicate that amounts of inhibitor large compared with those used by Tucker and Ball are necessary for complete inhibition of the carbonic anhydrase system. Additional support for the contention that carbonic anhydrase plays an essential role in the formation of the pancreatic juice comes from the experiments of Birnbaum

and Hollander (27) in which carbonic anhydrase inhibitor 6063 extensively inhibited the secretion of pancreatic juice.

The next question that arises is whether the HCO_3 transport is an active or passive process in the pancreatic secretion. Thus the anions could move across the cell membrane first as the result of an active process providing an electrochemical gradient to transport the Na ions up the concentration gradient. Conversely the cations could move actively, letting the anions follow. This point can be examined in the light of the intracellular concentrations of Cl and HCO_3. Ball and his colleagues have shown (16) that the intracellular concentration of these ions changes very little between rest and secretion. Therefore it does not seem likely that any passive mechanism could account for the changes in distribution of HCO_3 and Cl ions that accompany changes in rate of secretion. For the initial intracellular concentration ratio and the relative permeability of the cell wall to these two ions will set between them the ratio which these ions bear to one another as they appear in the pancreatic juice. So long as these two factors remain constant despite intensive secretion, the ratio between Cl concentration and HCO_3 concentration in the pancreatic juice should remain constant in any passive system. Since the observations of Hart and Thomas (7) in figure 2 show that the HCO_3 concentration is a function of secretion rate, it would follow that the anionic secretion involves an active process. Since this conclusion rests on the results of only two experiments of Ball *et al.* (16), and involves the assumption that the acinar cells are responsible for the secretion, it must be accepted with reserve. However, it would seem fruitful to carry out more experiments to clarify this point.

The conclusion that HCO_3 transport involves an active process does not necessarily mean that cation transport is passive. Cation transport may be either active or passive, and the HCO_3 transport system may or may not provide an electrochemical gradient which is utilized by the cations. In our present state of ignorance of the detailed mechanism of pancreatic secretion, it would seem more economical to implicate only a single process, the anionic transport, as an active process; however, such a concluson seeims premature until we learn more of the details of the total economy of the cell.

Next let us consider the alterations required in the hypothesis if it should develop that the acinar cells are not responsible for the electrolyte secretion. If such were the case, the arguments concerning the intracellular concentration of the various ions would disappear until such time as fresh individual determinations were made of the ionic concentration in the several types of cell. With the use of Birnbaum and Hollander's observation on the inhibition by inhibitor 6063, we could still assume that the CO_2 secretion mechanism were correct, and were an active process with Na and K ions following along in a purely passive fashion. The cell would be assumed freely permeable to cations until proved otherwise.

Note added in proof: Davies (32) in his doctoral dissertation has put forward a hitherto unpublished alternative hypothesis for the elaboration of the HCO_3 in the pancreatic juice. The secreting cells are considered to be similar to a "stomach working in reverse" and the primary reaction is considered to be the separation of H_2O into its component ions, H^+ and OH^-. The H ion is expressed into the blood stream in exchange for a Na ion. CO_2 enters the acinar cell in the gaseous phase, and there combines with the OH to form the HCO_3 which is released into the juice. These intracellular OH ions effectively remove intracellular CO_2 thus providing a concentration gradient which brings more CO_2 into the cell from the blood stream.

The important new feature of Davies' scheme is that the production of OH ions within the cell is assumed to be the primary process in the production of HCO_3. These ions are considered to be formed as the result of two specialized enzyme systems, one involving dehydrogenating enzymes for the transfer of H atoms, and the other involving the cytochrome system, or a cytochrome-like system for the transfer of electrons. Such a scheme would provide the essential link between the biochemical reactions within the cell and the physiological functions that the cell performs.

If it can indeed be shown experimentally that the intracellular production of OH is the primary pancreatic process, the questions of relative cell wall permeability become less important. However it is as necessary that there be a region of high local HCO_3 concentration near the cell wall on the duct side according to Davies' hypothesis, as it is necessary that there be a local region of high OH concentration using the scheme outlined above. Once it is accepted that OH ions are pro-

duced intracellularly, high local concentrations of both OH and HCO_3 would probably be found, and it would not seem unlikely that transport from the cell into the juice would take place according to both schemes.

ENERGY REQUIRED FOR PANCREATIC SECRETION

It has been pointed out from the studies of Chardon and Gross (10) that O_2 is required for pancreatic secretion, and hence metabolic energy is presumably utilized for the process. From the data of Still, Bennett, and Scott (14), it is possible to make an estimate of the amount of energy produced by the metabolism of the gland during secretion. By combining their figure of 0.062 cc. O_2 consumed/gm. wet tissue min. when the gland is actively secreting, with the observation of Tucker and Ball (15) that a 31 gm. pancreas can secrete 1.25 ml. juice/min., it is possible to calculate that 1.53 l. O_2, or 0.068 mols, are consumed by the gland in producing 1 liter juice. Kielley and Kielley (28) have shown that approximately 3 high energy phosphate bonds are formed/atom O consumed, and Oesper (29) has calculated that 10,500 cals. free energy can be obtained from the hydrolysis of a high energy phosphate bond. Combining these figures, we find that 63,000 cals. free energy are produced/mol of O_2 consumed, or that approximately 4300 cals. of free energy are produced by the gland for every liter of pancreatic juice which it makes.

In order to complete our energy balance sheet it is necessary to make a set of assumptions concerning the mechanism of juice formation. These have been chosen to give a maximal estimate. First let us assume that all the anion in the juice is HCO_3. This is admittedly an overestimate, but it leads to only a small error in increasing the maximal computed energy. Then let us assume that no energy is made available by the downhill processes, that is by the entrance of Na from the plasma to the cell, and the secretion of K from the cell to the juice. Likewise we neglect all processes concerned with Cl transport, which are all downhill on the assumptions already made, as well as any energy which may be made available from the transport of HCO_3 from the plasma into the cell. We will also assume that the activity coefficients for the various ions are identical inside and out. We need not take account of any potential which may exist inside the cell, since as many cations as anions pass the cell wall in each direction. With these assumptions, it is now possible to write the following equations for the free energy required to form one l. juice at 38°C.

for HCO_3

$$\Delta F = 0.154\ RT \ln \frac{154}{18.2} = 203 \qquad (3)$$

for Na

$$\Delta F = 0.154\ RT \ln \frac{154}{45} = 117 \qquad (4)$$

for K

$$\Delta F = 0.0048\ RT \ln \frac{105}{5.1} = 9 \qquad (5)$$

Summing these 3 equations, we arrive at a figure of 329 cal. required for the secretion of 1 liter juice at 38°C., which is a little less than 8 per cent of the 4,300 cals. available.

Ball (30) gives 133 calories as the energy required for the secretion of 1 liter of pancreatic juice as the result of a similar calculation in which he considers only the HCO_3 and Cl gradients from plasma to juice. This is close to the value of 162 calories which would be obtained using the figures in table 1, if we assume that the acinar cells are not involved in electrolyte secretion. In this case, we would consider the secreting cells to be cation permeable, and use, perforce, the plasma HCO_3 concentration of 28.3 rather than the intracellular one in *equation 3* above. The remaining small discrepancy between Ball's figures and the present ones arises from our assumption that all the anion in the pancreatic juice is HCO_3.

CONCLUSION

In conclusion it seems desirable to point out some of the lacunae in our present experimental knowledge of the pancreas, and to suggest certain avenues for further exploration. The primary point that is still unsettled is the question of what cells are secreting the juice. Closely coupled with this is the requirement for accurate intracellular analyses of the principal electrolyte pattern of the secreting cells. Another fascinating question is the validity of the assignment of the active role of anionic transport, and the subsequent elucidation of the method by which secretin accelerates the process. In this connection it would be interesting to study the action of secretin on other systems, particularly those which are rich in carbonic anhydrase.

It is a pleasure to acknowledge my indebtedness to all those who have been so helpful in discussing these questions with me. Dr. E. G. Ball has been primarily responsible for kindling my interest in the pancreatic problem, and I owe a particular debt to him, as well as to Dr. F. Hollander who has made many stimulating comments. I should like to add that the suggestions of the other members of the symposium have been of great assistance and that many of them appear in the present article. Mrs. Eleanor Huggins has carried out the cation analyses, and the work has been supported in part by the Atomic Energy Commission.

REFERENCES

1. Lim, R. K. S., S. M. Ling, A. C. Liu and I. C. Yuan. *Chinese J. Physiol.* 10: 475, 1936.
2. DeZilwa, L. A. E. *J. Physiol.* 31: 230, 1904.
3. Gilman, A. and G. R. Cowgill. *Am. J. Physiol.* 104: 476, 1933.
4. Johnston, C. G. and E. G. Ball. *J. Biol. Chem.* 86: 643, 1930.
5. Ball, E. G. *J. Biol. Chem.* 86: 433, 449, 1930.
6. Komarov, S. A., G. O. Langstroth and D. R. McRae. *Can. J. Res. D*, 17: 113, 1939.
7. Hart, W. M. and J. E. Thomas. *Federation Proc.* 4: 32, 1945; *Gastroenterol.* 4: 409, 1945.
8. Berg, B. N. and T. F. Zucker. *Proc. Soc. Exper. Biol. & Med.* 28: 724, 1930–31.
9. Babkin, B. P. *Die aussere Sekretion der Verdauungsdrusen.* Berlin: Springer, 1914, p. 310.
10. Chardon, G. and A. Gross. *Comptes rend. Soc. Biol.* 140: 1004, 1946.
11. Davies, R. E., A. A. Harper and I. F. S. Mackay. *Am. J. Physiol.* 157: 278, 1949.
12. Montgomery, M. L., G. E. Sheline and I. L. Chaikoff. *Am. J. Physiol.* 131: 578, 1940–41.
13. Oldfelt, C. O. *J. Physiol.* 102: 362, 1943.
14. Still, E. U., A. L. Bennett and V. B. Scott. *Am. J. Physiol.* 106: 509, 1933.
15. Tucker, H. F. and E. G. Ball. *J. Biol. Chem.* 139: 71, 1941.
16. Ball, E. G., H. F. Tucker, A. K. Solomon and B. Vennesland. *J. Biol. Chem.* 140: 119, 1941.
17. Eichelberger, L. and R. B. Richter. *J. Biol. Chem.* 154: 21, 1944.
18. Visscher, M. B. *Federation Proc.* 1: 246, 1942.
19. Fenn, W. O., D. M. Cobb, J. F. Manery and W. R. Bloor. *Am. J. Physiol.* 121: 595, 1938.
20. Hammarsten, E. and E. Jorpes. *Acta Med. Scand.* 68: 205, 1928.
21. Harned, H. S. and B. B. Owen. *The Physical Chemistry of Electrolytic Solutions.* Norfolk: Reinhold, 1950. (2nd ed.) Table (6-8-2), p. 172.
22. Sutra, G. *J. Chim. Physique* 43: 189, 1946.
23. Davenport, H. W. *Physiol. Rev.* 26: 560, 1946.
24. Roughton, F. J. W. *Physiol. Rev.* 15: 241, 1935; *Harvey Lectures* 39: 96, 1943–44.
25. Davies, R. E. and F. J. W. Roughton. *Biochem. J.* 42: 618, 1948.
26. Davies, R. E. *Biochem. J.* 42: 609, 1948.
27. Birnbaum, D. and F. Hollander. Private communication.
28. Kielley, W. W. and R. K. Kielley. *J. Biol. Chem.* 191: 485, 1951.
29. Oesper, P. *Phosphorus Metabolism.* (W. D. McElroy and B. Glass, Eds.) Vol. 1. Baltimore: Johns Hopkins Press, 1951, p. 523.
30. Ball, E. G. *Ann. N. Y. Acad. Sci.* 45: 363, 1944.
31. Davies, R. E. Personal communication.
32. Davies. R. E. Doctoral Thesis, Univ. of Sheffield, 1948.

AMERICAN PHYSIOLOGICAL SOCIETY

Panel Discussion: Interpretation and Significance of Alterations in Central Pulse Form[1]

Chairman: D. F. OPDYKE

INTRODUCTION

THE Council of the American Physiological Society inaugurated, at the New York meeting, a panel discussion, designed to fill an essential gap between a formal symposium and the short papers programmed in individual sections. Very often several groups of investigators are working on a single broad problem which by its nature presupposes a more or less specialized background. These workers may be loath to publish their results in a formal style while so many questions remain unanswered. They are just as reluctant to attempt a ten minute presentation of their views, omitting by necessity a survey of the background of the problem and the difficulties involved in the experimental attack upon it. The panel was therefore set up as a mechanism whereby representatives of the various working groups could meet on a semi-formal basis, and present a considerable body of experience and results for the purpose of stimulating criticism and suggestions from each other and from interested members of the audience.

In the organization of the present panel, it was felt that the discussion would be more profitable if we had present an unbiased referee, or moderator, who would make sure that questions were fairly put and adequately answered. It was also his province to outline, in an introduction, the problem under consideration. Dr. Opdyke agreed to undertake this unremunerative task, and became at once an enthusiastic member of the 'team.'

All members who participated in the panel would agree that it was a most profitable experience. Since the speakers, in their preparation, had had time to examine critically the logical strengths and weaknesses of their arguments and to frame their statements and questions in effective terms, more was accomplished than could ever be gained from impromptu corridor discussions. By permitting enough time to cover personal viewpoints, we are now much more aware of where our concepts are fundamentally the same, and where quite different. Needless to say, discussion on the points of controversy did not begin, or end, with the actual session. Perhaps the most important feature of such a panel was its effectiveness in lowering inter-laboratory barriers, and the promulgation of the mature viewpoint that the primary goal is the most rapid solution of the problem at hand, and not the attempted guarding of results until priority can be established in a published article.

The planning of the material to be presented, and the time schedule for such presentations, was done by the concerted effort of all panel members. From the viewpoint of a program most rewarding to the audience, mistakes were made in both categories, which can be laid mainly to inexperience in such planning. It was never proposed that any part of the program be published, since any such plans would have tended to formalize the presentations and rob the panel of much of its intrinsic worth. After the session, however, so many requests were received that

[1] New York City, April 16, 1952.

the individual contributions be written in the form of papers, that the panel members have agreed to prepare a single presentation of their views. In writing, they have felt free to omit any details which rightly belong in papers to be published later, and not to attempt any reconstruction of the actual discussion: the papers are, of course, colored by points raised in the discussion. These papers cannot capture the flavor of the meeting itself, and cannot, therefore, be taken as an index to what actually transpired at the session.

JOHN W. REMINGTON

GENESIS OF THE PRESSURE PULSE CONTOUR METHOD FOR CALCULATING CARDIAC STROKE INDEX[1]

DAVID F. OPDYKE

From the Merck Institute for Therapeutic Research, Rahway, New Jersey

THE need for a practical method of quantitatively assessing the stroke-by-stroke output of the heart is recognized by all investigators in the fields of cardiovascular physiology and pharmacology. Indeed, we may even risk the statement that the lack of such a method is the greatest single bottleneck impeding further progress in these fields. Since the need is so critical it is wise for us to consider seriously all newly proposed methods and to subject them to thorough evaluation.

The object of this panel discussion is to examine critically and debate the probability of quantitatively relating the contour of the aortic pressure pulse to the cardiac stroke index. The practical importance of this relation has already been stressed, but equally important is the information gained concerning aortic dynamics incident to attempts at putting the relation on a quantitative basis.

Attempts to quantitate stroke volume on the basis of calculations performed on the aortic pressure pulse are not a recent development. However, two groups within the American Physiological Society have recently been concerned with a new approach to the problem. It is only natural that differences of opinion have arisen. These have been aired privately with mutual benefit and stimulation. Since such interdepartmental discussions do not benefit other interested parties, the participants in this friendly warfare and their allies have been invited to compare, debate, and defend their views publicly.

It is the purpose of this introductory paper to trace briefly the history of the contour method, not only for the benefit of those who have had only a casual acquaintance with the problems involved, but also to put the subsequent remarks of our panel members in proper perspective. I shall attempt to establish in simple words the basic principles of the contour method, to show the mechanics of its application, and to summarize some of the results of its evaluation.

An aortic pressure pulse is a graphic representation of the pressure-time relation within the vessel, the relationship being dependent upon the rate of change of internal volume of the vessel, the distensibility of the vessel wall, the initial volume of the arterial bed, and the mechanical conditions under which the pulse is recorded. The rate of change of volume is, of course, the resultant of the rates of inflow and outflow. Thus, instantaneous volume has as two of its governing factors the cardiac stroke volume and the arteriolar drainage. Instantaneous pressure, on the other hand, depends upon the relationship between instantaneous volume and the distensibility of the vessel wall. These relations may be summarized by the classic equation:

$$E = \frac{\Delta p}{\Delta v} V \qquad (1)$$

[1] New York City, April 16, 1952.

where E = the volume elasticity coefficient, p = change in pressure, v = change in the volume and V = aortic diastolic capacity. Thus, if E and V are known and remain reasonably constant the relation between pressure and volume becomes calculable.

Our starting point in attempting to establish a quantitative relationship is an accurately recorded aortic pressure pulse. The characteristics of such pressure pulses are shown in figure 1. The wave typically has a steep anacrotic limb which gradually becomes rounded. The curve may fall slightly or continue to rise until the end of systole. The end of systole is sharply marked by an incisura followed by a dicrotic notch after which the curve declines slowly to the diastolic level.

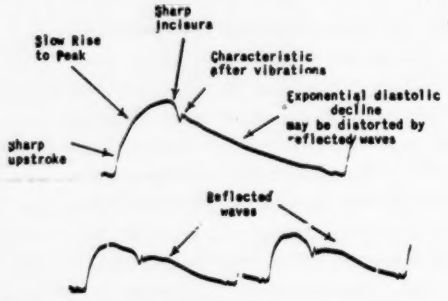


Fig. 1. CHARACTERISTICS of an accurately recorded aortic pressure pulse.

Although several embellishments may appear on the aortic curve in the form of reflected waves and natural frequency vibrations, time does not permit a discussion of these, nor can we discuss at this point the detailed instrumental requirements and technical precautions necessary in order to obtain an adequate curve. For our purpose we will take for granted that the central pulse is propagated intact out of the arterial tree with only minor changes due to frictional damping of the higher frequency components and the addition of reflected waves. It may be said, however, that the majority of the panel members agree that pressure pulses recorded by low frequency galvanometers (20–40 c.p.s.), or by transducers connected to impossible hydraulic systems, lack the detail, sharpness, and internal checks of accuracy necessary for precise calculation, statements to the contrary notwithstanding.

For the purpose of attempting to relate pressure change with volume change the aorta and its larger branches may be considered as a compression chamber or 'Windkessel'. This functions to buffer pressure changes and to provide storage space when inflow exceeds outflow. Theoretically, the pressure-volume relations of such a chamber should be predictable from *equation 1*. This Windkessel differs from a typical compression chamber, however, by virtue of the fact that the wall of the chamber is the distensible element. Furthermore, the aortic compression chamber departs from the classic concept in two other respects: *a*) The various parts of the aorta and its branches possess varying degrees of distensibility; and, *b*) the compression chamber is a branching tubular one rather than the classical sphere.

It is apparent that difficulty should be encountered in quantitating E and V. The chief difficulty in estimating the parameter of E is the fact that the modulus of elasticity is different in various regions of the arterial tree. Furthermore, the geographical limits of the Windkessel are not clearly defined. (e.g., does the resonating chamber include only the aorta or should the larger arteries and possibly the whole arterial bed be included?) It is understandable, then, why attempts have been made to arrive at an estimate of the mean elasticity of the whole compression chamber by methods which circumvent the necessity of quantitatively measuring the elasticity and volume of the various arterial beds.

Bramwell and Hill (1) attempted one short cut when they derived a relation between pulse wave velocity and overall elasticity. Their equation was:

$$Vp = K\sqrt{\Delta pV/\Delta V}. \qquad (2)$$

where Vp = pulse wave velocity, K = a constant.

Thus, pulse wave velocity was thought to be an index of distensibility. Utilizing this principle, both Broemser and Bazett (2, 3) attempted to quantitate the cardiac stroke volume from pulse pressure, pulse wave velocity, and estimations of aortic diastolic capacity, but with little success. Restudy of the problem by Hamilton, Remington and Dow (4) revealed the probable cause of the lack of quantitative agreement.

From studies on the transmission of generated pulses they established that the velocity of the foot of the pressure wave seemed to be measuring a 'reluctance to stretch,' a 'viscous action' or 'hysteresis' of the aortic wall which would be changed by differences in the rate of tension development. This was confirmed by subjecting rings to rapid and slow stretch and observing

'hysteresis loops' on the tension-length diagrams. After rapidly repeated fast stretches it was found that the tension-length diagram approached that of slow stretch. These results were found to be directly applicable to the dog aorta *in situ* and to aortic rings. It was concluded that pulse wave velocity could be equated with distensibility (and hence volume change) only if a suitable constant was introduced. Unfortunately, the constant could be applied only if diastolic pressure exceeded 70 mm. Hg. The conclusion was reached that pulse wave velocity could never be a quantitative measure of total aortic volume uptake in the living animals.

TABLE 1. CAPACITY PER SQUARE METER BODY SURFACE OF ARTERIAL TREE AND ITS SEVERAL PARTS AT DIFFERENT PRESSURES[1]

PRESSURE	ARCH	HEAD AND THORACIC	VISCERA AND ABDOMINAL	LEGS	TOTAL
mm. Hg	*cc.*	*cc.*	*cc.*	*cc.*	*cc.*
20	2.1	5.1	1.9	1.8	10.9
40	5.0	11.7	4.3	4.0	25.0
60	8.7	19.5	7.1	6.5	41.8
80	12.5	28.2	9.7	8.7	59.1
100	16.2	36.6	11.7	10.7	75.2
120	19.1	42.9	13.7	12.5	88.2
140	21.2	48.1	15.3	14.0	98.6
160	22.9	51.9	16.6	15.2	106.6
180	24.0	54.0	17.5	16.0	111.5
200	24.9	55.8	18.3	16.7	115.7
220	25.7	57.3	19.0	17.3	119.3

[1] After HAMILTON AND REMINGTON (6).

From studies of putrefying aortas Remington, Hamilton and Dow (5) confirmed and extended earlier work showing that the distensibility of the aorta was conditioned by three anatomical systems, muscle, elastic tissue and collagenous fibers. Each system had its own inflection point on the distensibility curve. This observation, together with data on slow stretch of aortic rings and measurements of aortic capacity from dead dogs (and human aortas?) led Hamilton and Remington (6) to decide that little correlation existed between aortic capacity and absolute distensibility. They found the relation $\frac{\Delta V}{V \Delta P}$ (relative distensibility) to be much more variable than the absolute distensibility $\frac{\Delta p}{\Delta v}$. Thus, they ignored absolute aortic capacity, assuming it to be a quantity which would vary with body surface area, and therefore correctable, and proceeded to compile tables of volume uptake for each arterial bed based only on $\frac{\Delta p}{\Delta v}$. The original constructed table, and a table of transmission times for the propagated pressure pulse, are reproduced in tables 1 and 2. This short cut, circumventing the necessity of quantitating absolute aortic capacity (V), represents the point of departure from previous attempts to quantitate the aortic pressure pulse in terms of stroke volume.

Having a pressure-volume table, Remington and Hamilton (7, 8) thought it feasible to attempt

TABLE 2. PULSE WAVE TRANSMISSION TIMES TO PARTS OF ARTERIAL TREE AT VARIOUS DIASTOLIC PRESSURES FOR DOG OF 15 KG.[1]

PRESSURE	ESTIMATED TRANSMISSION TIMES			
	FOR UPTAKE			FOR DRAINAGE
	Head and Thoracic (Th)	Visceral and Abdominal (Tv)	Leg (Tl)	Wted. Av. (Tw)
mm. Hg	*msec.*	*msec.*	*msec.*	*msec.*
20	59	110	170	129
40	48	89	136	101
60	39	75	112	84
80	33	64	94	72
100	28	55	81	62
120	24	47	70	53
140	20	40	59	46
160	18	34	50	39
180	16	28	42	33

[1] After HAMILTON AND REMINGTON

the construction of a derived cardiac ejection curve from an aortic pressure pulse, the pressure-volume table, known transmission times, and a calculation of systolic drainage.

The progressive systolic uptake U of each division of the aorta (i.e., the increase of aortic volume during systole) was calculated per unit of time, the volumes being calculated from increase of pressure and transmission time. Systolic drainage was calculated on the basis that the volume outflow during the ejection period was proportional to the pressure over 20 mm. Hg and that the aortic uptake during ejection was equal to the diastolic outflow. Thus, by estimating systolic area as effective pressure times duration and diastolic area likewise, total systolic drainage could be calculated by the equation:

$$Sd = \frac{Sa}{Da} \times U \qquad (3)$$

where Sd = systolic drainage, Sa and Da = systolic and diastolic drainage areas respectively, and U = arterial uptake. The calculated systolic drainage was then apportioned over the ejection period on the basis of the proportion of pressure rise per 10 millisecond interval. These volumes were added to the corresponding systolic uptake and the volumes of each time interval summated to give the cardiac ejection curve.

rately, were consolidated into four great regions, arch, head, visceral and leg, all lying, in time, about the same distance from the heart and each region treated as a unit. Secondly, rather than quantitate the stroke volume from a constructed ejection curve they elected to calculate aortic uptake at a single point on the pressure pulse (closure of the semilunar valve) and add to this figure the calculated arteriolar drainage during

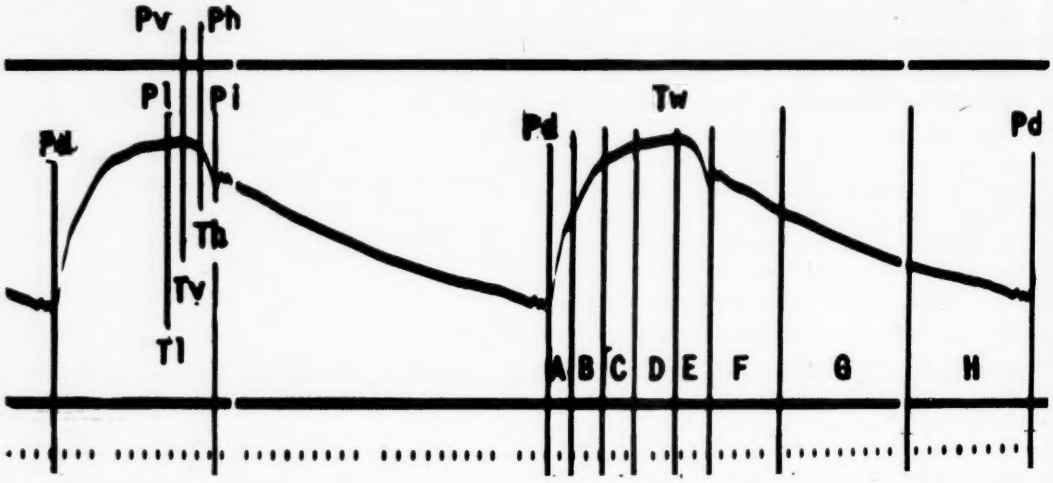


Fig. 2. STEPS IN CALCULATION of stroke index from the contour of an aortic pressure pulse. *Left hand curve*, location of pressure points for determining arterial uptake. *Right hand curve*, calculation of systolic and diastolic drainage areas. Sum of *areas A, B, C*, and *D* equal systolic drainage area. Sum of *areas E, F, G*, and *H* equal diastolic drainage area (see text).

The correspondence of contours between the derived curve and cardiometer curve was good, but due to the unknown effect of changes in aortic tone it was emphasized that the stroke volumes calculated from such derived curves lacked accuracy.

In summary, estimations of stroke volume based on calculations employing pulse wave velocity have been found inaccurate. Remington and Hamilton have explored a new approach. This is by way of appraising the volume-pressure relations of the various parts of the aortic system, and deriving a cardiac ejection curve which satisfied the contour of the aortic pressure pulse. It remains for our panel to discuss the validity of the assumptions on which this derivation is based.

Proceeding on the basis of the excellent quantitative fit of the derived ejection curve with the arterial uptake curve, Hamilton and Remington (6) proceeded to simplify the laborious calculations in an attempt to establish a simple, practical method of calculating stroke volume. First, the various arterial beds, previously treated separately, were consolidated into four great regions, arch, head, visceral and leg, all lying, in time, about the same distance from the heart and each region treated as a unit. Secondly, rather than quantitate the stroke volume from a constructed ejection curve they elected to calculate aortic uptake at a single point on the pressure pulse (closure of the semilunar valve) and add to this figure the calculated arteriolar drainage during systole, the total being equal to stroke volume according to the equation:

$$SV = U + Sd \qquad (3)$$

where SV = stroke volume, U = arterial uptake, and Sd = systolic drainage.

The steps in the calculation are illustrated on

TABLE 3. SUMMARY OF EXPERIENCES WITH HAMILTON AND REMINGTON AORTIC PRESSURE PULSE CONTOUR METHOD OF EVALUATING CARDIAC STROKE INDEX

INVESTIGATOR	CHECKED AGAINST	AV. DIFF. %	RANGE %
Hamilton and Remington	Dye Injection Direct Fick	±8.2	−13–+35
Remington, Hamilton, Wheeler and Hamilton	Dye Injection Direct Fick	±10	−50–+45
Huggins, Handler and LaForge	Direct Fick	±12.2	−15.5–+23
Huggins, Smith and Sinclair	Direct Fick (open chest)	±12.5	−13.7–+50
Longino and Gregg	Rotameter (open chest)	±13.1	−40–+37
Duomarco, Dillon and Wiggers	Differential Flowmeter (open chest)	±142	−14.0–+374

figure 2 and are performed as follows: *1*) The total aortic uptake at the end of ejection is calculated for each of the four divisions of the aorta. This involves laying back from the incisura the transmission times (from table 2) from the arch to the head and thorax (*Th*), abdominal and visceral (*Tv*), and leg beds (*Tl*), as shown on the left hand curve of figure 2, in order to determine the simultaneous pressures in these parts of the arterial system. The pressure at these respective points is referred to table 1 and the uptake for each division obtained. *2*) The uptakes for all divisions are totaled to give *U*. *3*) Calculation of systolic drainage: this involves the determination of the ratio between systolic and diastolic drainage areas. *a*) The diastolic drainage area is calculated as mean diastolic pressure times duration as shown on the right hand curve of figure 2. *Tw* is derived from average transmission time to the various arteriolar beds weighted by their relative drainage. Adding the area bounded by *Tw* and the incisura to the diastolic drainage area takes into account that systolic drainage begins when the pulse wave reaches the periphery and ends at the time of the closure of the aortic valves, all drainage thereafter being at the expense of blood already accounted for. *b*) The systolic area is calculated as the mean systolic pressure times the systolic drainage time. *c*) The systolic drainage is then calculated by substitution in *equation 3* above and the drainage figure added to *U* to give the stroke volume.

It should be emphasized that the stroke volume figure so derived is stroke index and must be multiplied by the body surface area to obtain actual stroke volume.

Comparisons of calculated stroke volumes with stroke volumes obtained from direct Ficks, dye-injection or rotameter measurements have been made under a variety of conditions by several groups of investigators. Their findings are summarized in table 3. The average differences for all but one investigation are between ±8.0 and ±13.0 per cent, which is very nice. But the range of individual variation is from −50 to +45 per cent. The randomness of the individual variation is a source of concern and the subject of the major part of this discussion.

REFERENCES

1. Bramwell, J. C. and A. V. Hill. *Proc. Roy. Soc. London* 93: 298, 1922.
2. Broemser, P. and O. F. Ranke. *Ztschr. f. Kreislaufforsch.* 25: 11, 1933.
3. Bazett, H. C., F. S. Cotton, L. B. La Place and J. C. Scott. *Am. J. Physiol.* 113: 312, 1935.
4. Hamilton, W. F., J. W. Remington and P. Dow. *Am. J. Physiol.* 144: 521, 1945.
5. Remington, J. W., W. F. Hamilton and P. Dow. *Am. J. Physiol.* 144: 536, 1945.
6. Hamilton, W. F. and J. W. Remington. *Am. J. Physiol.* 148: 14, 1947.
7. Remington, J. W. and W. F. Hamilton. *Am. J. Physiol.* 144: 546, 1945.
8. Remington, J. W. and W. F. Hamilton. *Am. J. Physiol.* 148: 25, 1947.

FACTORS DETERMINING THE CONTOUR OF PRESSURE PULSES RECORDED FROM THE AORTA

ROBERT S. ALEXANDER

From the Department of Physiology, Western Reserve University School of Medicine, Cleveland, Ohio

PRESSURE pulses from the central aorta may be readily recorded from the experimental animal and, by the use of polyethylene catheters (1), are now obtainable in the human patient. If one takes the necessary precautions to *a*) acquire adequate recording instruments, *b*) develop adequate techniques of using those instruments so as to avoid artifacts, and *c*) record at sufficient tracing speeds to demonstrate clearly the details of pulse contour, accurate recordings of pressure pulses in the aorta appear to present a wealth of information regarding the circulatory status of the animal or patient under observation. The empirical value which an experienced investigator can realize from the inferences he draws from such aortic pressure pulses (2, 3) encourages attempts to devise an objective method of quantitating some of the circulatory variables involved, of which the most important would be the stroke volume output of the heart.

Since empirical methods of such quantitation have been notorious for their failures, it becomes important to inquire as to whether our understanding of the factors determining pulse contour is adequate to justify a rational solution to the problems of quantitation. It is the purpose of this presentation to examine such pulses from a qualitative point of view in order to define the variants which must be accounted for in a valid quantitation. The material to be presented has been selected from a collection of several thousand pulse tracings which have been obtained from studies on anesthetized dogs under a wide variety of experimental conditions, employing optical manometer methods as have been described elsewhere (4).

At the outset it is important to set up a working hypothesis as to what is represented by the pressure changes in the aorta during a cardiac cycle. As a first approximation, the aortic system has been classically regarded as an elastic compression chamber or 'Windkessel,' which can be thought of as an elastically distensible sphere with cardiac ejection entering at one pole and drainage to the peripheral vascular bed leaving from the opposite pole. It should be understood that this is only a first approximation which will be subject to some important modifications, including especially those discussed by Remington (5). On this basis, four factors immediately become of major importance in determining the pressure and the increments of pressure change from moment to moment: *1*) the modulus of elasticity of the walls of the Windkessel, *2*) the volume of blood within the Windkessel, *3*) increments in volume derived from cardiac ejection, *4*) decrements in volume due to peripheral drainage. The first factor would remain constant and predictable unless one accepts the possibility of vasomotor effects on the aortic Windkessel, as will be discussed later. On the same basis, the volume within the Windkessel at the onset of systole should be a function of the diastolic pressure. Increments in volume derived from cardiac ejection will be determined by the precise manner in which the myocardium contracts during systole, which in turn will be dependent to a major extent upon the classical Starling principles relating the vigor of contraction to the degree of filling of the ventricles during diastole. The peripheral drainage will follow the Poiseuille principles of blood flow, determined essentially by the pressure within the Windkessel and the resistance to peripheral flow, the latter being dominated by vasomotor influences.

With these basic concepts in mind, let us examine some of the more typical changes in aortic pulse contour to determine how adequate these first approximations appear to be for explaining the observed pulse forms.

EXAMINATION OF BASIC PULSE FORMS

A. Influence of changes in venous return to the heart. Perhaps the simplest dynamic alteration in aortic pulse form is that associated with an acute decrease in venous return to the heart with a consequent reduction in the stroke volume ejected into the aortic Windkessel. This may be

readily demonstrated in the animal with open chest by suddenly constricting the pulmonary artery and thereby producing an acute reduction in venous return to the left ventricle, revealing pulse changes as shown in figure 1. By scanning this recording from left to right, a number of alterations in the detailed contours are noted. The initial systolic upstroke of the pulse becomes reduced in both amplitude and rate. This would be explained as the consequence of a smaller and less vigorous contraction of the myocardium due to reduced diastolic filling. The systolic peak, which is well rounded on the initial pulse, tends to flatten and slump to a relatively lower incisura. This correlates with cardiometer evidence that a poorly filled heart tends to complete its active ejection during the earlier part of systole (6). The diastolic limb becomes flatter as the stroke volume and pressure level fall. This would relate to the fact that peripheral drainage from the Windkessel is dependent upon pressure, and drainage will therefore be reduced as pressure falls. Supplementing this, reflex vasoconstriction would further reduce peripheral drainage when sufficient hypotension and time had elapsed for that reflex to exert a significant influence.

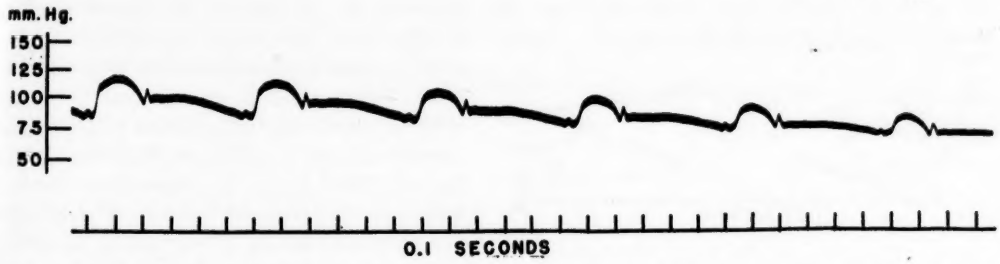


Fig. 1. PRESSURE PULSES recorded from the aortic arch of an anesthetized dog; pulmonary conus partially occluded at the start of the record.

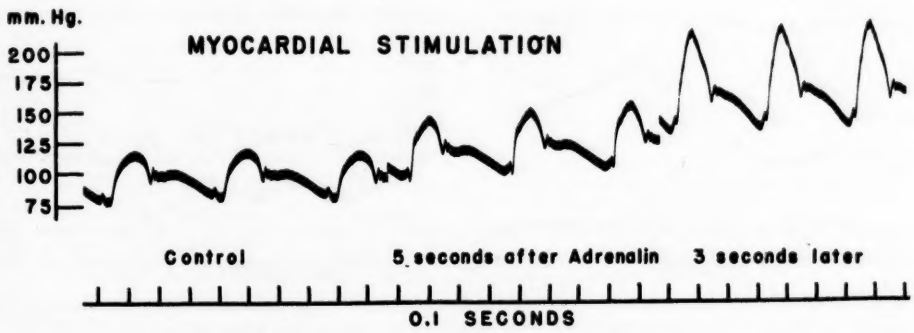


Fig. 2

B. Changes in myocardial contractility. Figure 2 illustrates the acute changes associated with myocardial stimulation produced by injecting 1 mg. Adrenalin directly into the superior vena cava. It may be assumed that the initial changes observed under these conditions are those attributable dominantly to myocardial stimulation rather than to peripheral systemic vasoconstriction. The first change noted is a marked sharpening of the initial systolic upstroke, signifying a much more vigorous contraction of the myocardium. Accompanying this there is also evidence of an abbreviated isometric contraction phase of the heart. Corresponding roughly with the initiation of isometric contraction of the ventricle, the aortic pulse exhibits a characteristic irregularity in late diastole which appears in the control pulse of figure 2 as a discrete joggle to the left of and clearly dissociated from the upstroke due to systolic ejection. With myocardial stimulation this joggle becomes merged with the foot of the pulse rise as the Adrenalin effect becomes manifest. There is also a marked peaking of the pulse to a sharp and early systolic peak. All of these alterations are compatible with a very vigorous contraction of the myocardium which very rapidly builds up sufficient pressure to eject blood into the aorta and in turn expels its stroke

volume vigorously and rapidly during the early portion of an abbreviated aortic systole. The validity of this interpretation of the pressure contour with myocardial stimulation has been directly confirmed by cardiometer studies of the ejection process under these conditions (6). The sharp pre-incisural fall in pressure together with the marked steepening of the diastolic limb indicates an increase in peripheral drainage due to the high systolic peak of pressure. Since we may assume that there has not been any significant degree of peripheral vasoconstriction at this time,

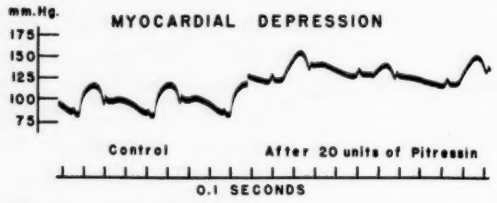


Fig. 3

out during the remainder of systole. It should be noted that the depressed heart exhibits a pulsus alternans, as is quite commonly observed under these conditions, and that the weak beat shows an even greater exaggeration of the altered contour than does the strong beat. The tendency of the diastolic limbs to flatten somewhat under the influence of Pitressin presumably relates to some degree of peripheral vasoconstriction which acts to reduce peripheral drainage from the Windkessel.

C. Vasomotor changes altering peripheral drainage. It is difficult to induce vasomotor changes in the periphery without simultaneously altering other factors contributing to pulse contour. The recordings on the left of figure 4 combine myocardial stimulation with peripheral vasoconstriction. In contrast to the recording of figure 2 which showed the immediate effects of a large dose of Adrenalin, the recording in figure 4 was obtained during a continuous intravenous

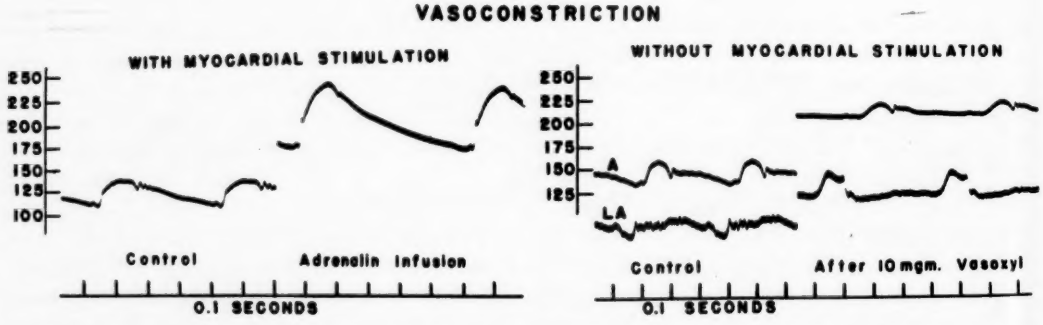


Fig. 4. *Pair of curves on left:* AORTIC PULSE TRACINGS as described in text; *pair of curves on right:* aortic pulse tracing (*A*) above and left atrial pressure tracing (*LA*) below as described in text. Pressure scales refer to aortic pulse only.

and since the passive effect of a pressure rise is to distend the peripheral vessels and thereby effectively decrease the resistance to flow (7), rapid peripheral runoff with a consequent sharp drop in pressure is to be expected.

Figure 3 illustrates the converse of the above, where the myocardium was depressed by administering a massive dose of Pitressin. In this case the isometric joggle preceding systolic ejection is separated from the ejection phase even further than in the control. The slope of the upstroke is more gradual, rising slowly but persistently to a late systolic peak. This represents the sluggish contraction of a depressed myocardium which requires considerable time to develop enough pressure to eject blood into the aorta, and then relatively slowly wrings the blood

infusion of Adrenalin which had been maintained for about 2 minutes before this record was obtained. The pulse shows an abbreviated isometric contraction phase and rapid initial upstroke associated with the myocardial stimulation, as has been described above. In this instance, however, the increase in peripheral resistance serves to greatly elevate the pressure level so as to impose a marked increase in the output load against which the ventricle must pump. Accordingly, the heart does not empty early as in figure 2, but shows a maintained ejection to a late systolic peak and a high incisura. The slope of the diastolic limb during Adrenalin infusion is steeper than that in the control because of the much higher pressure level of the incisura and consequent increase in the driving force for peripheral

drainage. Actual calculations from such diastolic limbs, however, reveal that the rate of diastolic fall in pressure is not as great as would be expected on the basis of the very large rise in incisural pressure, as might be guessed by comparing the diastolic limb of the Adrenalin infusion in figure 4 with that observed in figure 2. This failure to observe a more rapid fall in pressure during the Adrenalin infusion indicates significant vasoconstriction in the periphery which prevents the blood from draining from the Windkessel as rapidly as would be predicted on the basis of pressure alone.

The pair of curves[1] on the right of figure 4 represent in contrast the effect of vasoconstriction in the absence of myocardial stimulation and, in fact, with some relative if not absolute myocardial depression. The aortic pulse at the top of the recording demonstrates relative stagnation of blood within the Windkessel, as indicated by the flat diastolic limbs in spite of the incisural pressure of 212 mm. Hg. Because of the very high diastolic pressure, there is a prolonged isometric contraction phase before the ventricle can develop enough pressure to eject blood into the aorta. This is not only reflected in the position of the isometric joggle on the aortic pulse tracing, but is also demonstrated by the sharp deflection on the left atrial tracing (*LA*) which signifies the onset of isometric contraction of the ventricles. Once ejection is initiated the heart must still perform a great deal of work to eject blood at this pressure level, and hence the upstroke of ejection is gradual, terminating in the late systolic peak which is characteristic of vasocontriction.

Figure 5 illustrates the changes associated with vasodilation. Adenylic acid was chosen as the vasodilator because of its minimal effect on the heart. Since the heart is well filled and has a lower aortic pressure to work against, its initial ejection is rapid and the systolic peak occurs early, as shown on the upper tracing of figure 5. Because of dilation of the peripheral vascular bed, the peripheral runoff of blood from the Windkessel starts to exceed the amount being injected by the heart in midsystole, and hence the pressure starts to drop well in advance of the incisura and continues to drop steeply throughout diastole. The full significance of this diastolic pressure drop can be appreciated by referring back to figure 1, in which it was shown that a drop in pressure in itself tended to reduce peripheral drainage and hence flatten the diastolic limb. The observation in figure 5 of a steep diastolic limb in spite of the lowered pressure level would indicate a marked reduction in the resistance factor influencing drainage from the Windkessel.

The above description of figure 5 ignores the very obvious spike which occurs at the top of the anacrotic upstroke of the aortic pulse. At the moment this will simply be identified as a 'water-hammer' manifestation whose genesis will be discussed in more detail below. It should be noted at this time, however, that such pulse abnormalities are frequently evidence of inadequate apparatus or faulty technique. Evidence that this water-hammer spike is not a recording artifact is given by the middle tracing in figure 5 of the pressure pulse in the abdominal aorta. This pulse shows a comparable irregularity at the anacrotic crest although its magnitude has been considerably dampened, attesting that some genuine vibration must have occurred within the aortic Windkessel to account for its appearance in two independent manometer systems.

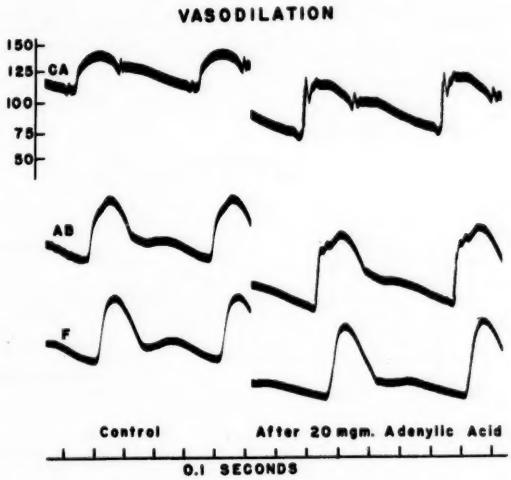


Fig. 5. PRESSURE PULSES recorded from the aortic arch (*CA*), the abdominal aorta at the level of the renal artery (*AB*), and a side branch of the femoral artery (*F*).

D. Multiple actions. In the above illustrations the effects of a single variable have been emphasized. It must be borne in mind that under the majority of physiological situations a number of variables may alter simultaneously. It is therefore instructive to consider such a combination of changes as illustrated by the response to hemorrhage and hypotension illustrated in figure 6.

[1] We are indebted to Dr. Jay Ankeney for these tracings which demonstrate by the use of Vasoxyl (Burroughs Wellcome and Co.) the most severe example of vasoconstriction that we have observed.

The upper tracings indicate the changes observed in the aortic pulse. The first three pulses illustrate the immediate changes in response to a stepwise hemorrhage procedure (9), and are dominated by evidence of reduced stroke volume, as represented by a slower initial ejection, a slower initial upstroke, and a flattened systolic summit dropping down to a lower incisura. The flattened diastolic limb may be explained in part by the lower pressure level resulting in less peripheral drainage. Increased peripheral resistance produced reflexly could also be involved in reducing runoff and thus flattening the diastolic limb, correlating with the reflex acceleration of the heart as has become manifest in the third tracing. The subsequent tracings were obtained at intervals of maintained hypotension over a 2-hour period. It should be noted that in the fourth and fifth tracing of figure 6 there is evidence of further cardiac stimulation, in the acceleration in heart rate and some abbreviation of systolic duration. Associated with this myocardial stimulation is a more rapid initial upstroke terminating in a spike vibration which we would again classify as a water-hammer effect. This type of pulse with a sharp early systolic peak, a low incisura, and flat diastolic limbs is quite characteristic of these conditions. The last tracing in figure 6, with a falling pressure in the absence of further bleeding and a pressure level which has reached critically low levels, is associated with incipient myocardial failure. Under these conditions the evidence of myocardial depression is noted by the replacement of the rapid upstroke, observed in the previous record, with a very sluggish upstroke in spite of the fact that the load against which the heart must work is now minimal.

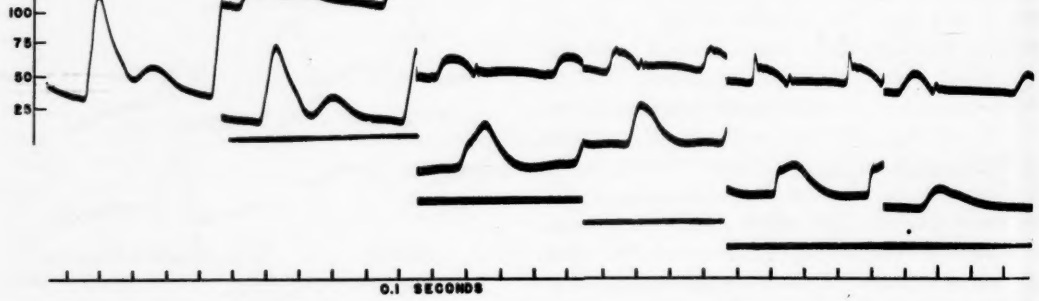


Fig. 6. Pressure pulses recorded from the aortic arch (*above*) and the femoral artery (*below*) during hemorrhage and maintained hemorrhagic hypotension. Pressure scale at left refers to aortic pulse only.

In the foregoing analysis an attempt has been made to explain some of the common alterations in aortic pulse contour in terms of the simple Windkessel hypothesis stated at the outset. This hypothesis appears adequate to explain the gross features of the pulses presented. It must have been obvious to the critical reader, however, that certain minor features of the pulse contours were being ignored and certain other aspects inadequately explained. If one is interested in using the aortic pressure pulse for quantitation it becomes important to define the minor variants as well as the major variants if one is to have assurance that his quantitation will prove valid. We must, therefore, proceed to modify the simpler working hypothesis to account for further factors which appear to contribute to aortic pulse contour.

DYNAMIC CONTRIBUTIONS TO PULSE CONTOUR DERIVED FROM AORTIC SYSTEM

A. The standing wave. A characteristic feature of most aortic pulses is an undulation in the pressure starting in midsystole and persisting into diastole. As usually encountered, this is represented by a pre-incisural slump in pressure and a post-incisural hump in the pulse contour. A somewhat exaggerated example of this is shown in the first cycle on the left of figure 7, although similar undulations may be noted in many of the preceding figures, including especially figures 2 and 3. The fact that these undulations are dependent upon phenomena occurring within the Windkessel itself is illustrated in the right-hand portion of figure 7. At the point indicated by the

arrow the abdominal aorta was occluded at the bifurcation of the iliacs. It will be observed that there is immediately a marked change in these undulatory changes in pressure preceding and following the incisura.

These undulatory changes in pressure have been subjected to considerable study. A complete discussion of this aspect of the aortic pulse contour is not pertinent to the present problem, but its major features should be appreciated. Hamilton and Dow (10), following up an earlier suggestion by Otto Frank, demonstrated that these undulations were produced by an over-all oscillation of the entire blood column in the aorta, oscillating about a node located near the diaphragm. Since it is important to distinguish the pulse wave itself, which is transmitted progressively along the wall of the vessels, from these

Fig. 7. PRESSURE PULSES recorded as in figure 5; arrow indicates point of occluding aorta at the bifurcation of the iliacs. Time lines below indicate 0.1 sec. (from *Am. J. Physiol.* 158: 297, 1949).

to-and-fro surges of the entire aortic blood column, the pressure oscillations were described as the 'aortic standing wave.' In previous publications (4, 11) we have confirmed and somewhat extended the concept of the standing wave as presented by the Georgia group, but we have not been satisfied that the actual forces involved in the genesis of the standing wave have been adequately defined. Recent work, however, permits a more specific explanation of its genesis (12), as will now be reviewed briefly.

Measurements indicate that the standing wave is first manifest at a point in time which corresponds to the round-trip pulse transmission time from the root of the aorta to about the level of the diaphragm, or to a point corresponding approximately with the node of the standing wave as determined by Hamilton and Dow. This suggests that some dynamic event occurs when the pulse wave reaches this point which starts the entire system oscillating. Reference to the anatomical relationships of the aorta at this level, as shown in figure 8, indicates that after traversing a significant length of the thoracic aorta which is devoid of any major branches, the advancing pulse wave suddenly encounters a group of very large arteries represented by the coeliac and superior mesenteric followed shortly by the two renal arteries. Measurements reveal that the cross sectional area of these vessels greatly exceeds that of the thoracic aorta, which would indicate a sudden lowering of resistance at this point. Basic principles of fluid mechanics

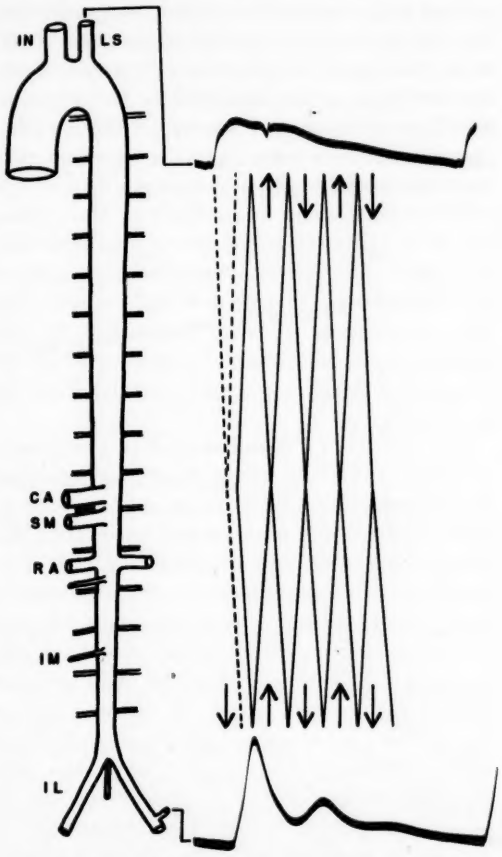


Fig. 8. SCHEMATIC DIAGRAM of the standing wave pressure undulations related to a scale drawing of the aorta and its major branches (*left*) and to actual pressure pulse recordings (*right*) obtained from the aortic arch and the femoral artery. For details refer to text.

predict that when such an advancing pressure wave reaches a point of lowered resistance, a forward surge will result. This surge will summate with the transmitted wave and result in a large pressure wave entering the femoral system. Since the femoral bed represents increased resistance, the surge toward this bed will now rebound back toward the aortic arch. The surge toward the femoral bed initiated by the low resistance at the

diaphragm will be represented by a negative pressure wave in the pulse recorded at the arch (accounting for the pre-incisural slump), occurring simultaneously with the augmentation of the systolic peak in the femoral artery. The rebound of the blood column from the femoral bed will produce the negative dicrotic dip in the femoral pulse and a positive pressure wave in the central aorta (represented by the post-incisural hump). The right-hand side of figure 8 graphically schematizes this process, the dashed lines indicating the initiation of this process by the foot of the pulse wave, and the solid lines representing the standing wave peaks produced by the surges of blood represented by arrows. Although this process originates primarily as a reflection phenomenon, it should not be confused with a simple reflected wave, since, as indicated by the arrows, a surge in a given direction occurs simultaneously throughout the system. Supplementary tests of this hypothesis (12) appear to substantiate this basic mechanism as being responsible for the standing wave undulations in pressure which are manifest in pressure tracings recorded from the aortic system.

Recognition of the standing wave phenomena prompts some modifications of our previous statements explaining alterations in pulse contour in terms of the Windkessel concept alone. Thus, the late systolic peak and high incisura of the pulses observed with vasoconstriction (fig. 4) appear to result to a significant degree from constriction of the splanchnic bed. This would tend to negate the normally low resistance at this point and thus reduce the forward surge of blood which is responsible for the normal pre-incisural slump in aortic pressure. Conversely, the steep pre-incisural slump observed in vasodilation (fig. 5) would be explained in part by an exaggeration of the aortic standing wave due to an increased surge when the pressure wave reaches the dilated splanchnic beds. It is interesting to speculate as to how much of the late systolic drop in pressure observed with myocardial stimulation in figure 2 may also be due to standing wave phenomena, although it is difficult to dissociate this factor from the other dynamic events occurring simultaneously.

B. Kinetic factors associated with the ejection process. There remains to describe the phenomenon, previously noted in figures 5 and 6, which I have identified as a 'water-hammer.' Such a 'spike' or 'ear' at the anacrotic crest has often been encountered by investigators in the field. Such a manifestation is most commonly the result of errors in recording technique, but in certain instances it appears to be a genuine manifestation of some process occurring in the central aorta. Although the suggestion that such spikes represent some sort of water-hammer mechanism is not new, there has been to my knowledge no clear explanation as to just how they originate. It may be of interest to recount, therefore, that both Peterson (13) and I appear to have stumbled upon essentially the same explanation quite independently and virtually simultaneously on the basis of analagous manifestations which we both happened to observe when artificial injections are made into the venous system. Although my apparatus was not adequate for use in aortic injections, Peterson has been able to extend his observations to the arterial system to show that the analogy actually applies.

The observations I made, which should be compared with those reported by Peterson in the accompanying paper (13) were obtained by measuring the venous pressure of an isolated intestinal loop in which the arterial and venous circulations were momentarily interrupted while an injection of blood was made into the venous end. The injection was accomplished by means of a syringe coupled through a gear train to a motor drive so as to permit a uniform injection at the rate of 1 ml/sec. Bearing in mind the fact that this technique involved a small, closed, vascular segment, simple Windkessel theory would predict a progressive rise in pressure determined by the distensibility characteristics of the vascular bed. Actual recordings, as illustrated in figure 9, conform with this general prediction except for the occurrence of a very prominent spike and some secondary irregularities at the moment that the injection was started. The analogy between this initial spike and the water-hammer spikes which have previously been observed in aortic pulses is readily apparent.

It is now generally conceded that the kinetic factor in cardiac ejection is of relatively trivial consequence in terms of energy or pressure equivalents when distributed over the entire cardiac cycle. Recognition of this does not deny the possibility that a momentary deceleration of the ejected blood would result in a transient pressure manifestation of some significance. Such a process, indeed, appears to be responsible for the pressure artifact which one observes when recording from a length of blood vessel which is occluded distally (14). Moreover, in view of the

pressure tracings reported here, one may justifiably ask: Are there any circumstances in the unoccluded aortic system where there might be produced a water-hammer phenomenon, defined as the pressure manifestation of a sudden deceleration of the blood ejected by the heart?

This question would appear to be readily answered by comparing the rate of cardiac ejection with the ability of the aortic system to move along this increment of blood by the forward transmission of the pressure-volume increment as represented by the pulse wave. For example, consider the case of a normal dog with a diastolic pressure of 100 mm. Hg and a pulse wave velocity in the ascending aorta of 4 meters per second. To serve as a standard basis for comparing cardiac ejection rates we may take the interval of 0.02 second as a period during which ejection occurs at an essentially maximal rate. A normal dog will expel approximately one third of its total stroke volume of about 15 cc. in this interval, or about 5 cc. Since the cross sectional area of the aorta at this pressure would be 2.6 cm.^{2}, the linear velocity of ejection must be 1.9 cm. per 0.02 second or ca. 1 meter per second. Obviously, a pulse being transmitted at the rate of 4 meters per second is entirely adequate to accommodate this blood ejected by the heart.

On the other hand, assume the diastolic pressure drops to 40 mm. and the heart has been stimulated reflexly to eject more vigorously. Under such circumstances a heart may eject at least ⅔ of its stroke volume in 0.02 second. In addition, the drop in pressure will have resulted in a narrowing of the elastic aorta, so that its cross sectional area will now be only 1.4 cm.^{2}. If, for the sake of illustration, we use the same stroke volume as before, then 10 cc. of blood must be ejected in 0.02 second through this orifice, requiring a linear velocity of 3.6 meters per second. However, according to the Bramwell-Hill equation, the pulse wave velocity at this pressure would be reduced to 2.4 meters per second in the ascending aorta. We thus have blood being ejected with a velocity of 3.6 meters per second into an aorta which is capable of advancing this increment at the rate of only 2.4 meters per second. This must inevitably result in a sudden deceleration of the ejected blood and consequently a water-hammer phenomenon, giving an inordinate pressure rise the moment ejection starts. In other words, the sharp rise in pressure does not represent a significant increment in volume in the Windkessel as a whole, but merely a very small volume of blood which, since it is obstructed from advancing, must create an inordinate pressure-volume increment in a local area at the base of the aorta.

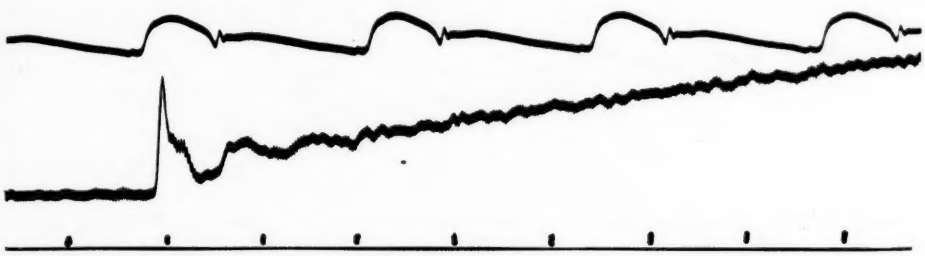


Fig. 9. *Upper record:* AORTIC PRESSURE PULSE. *Lower record:* Venous pressure recorded from isolated intestinal loop. The arterial and venous supply to this loop were suddenly occluded just before obtaining this record. The deflection of the venous pressure tracing coincides with the initiation of an injection into the vein of the loop at the uniform rate of 0.1 ml/sec. as described in text. Time marks indicate 0.2 sec.

On the basis of such calculations, employing data of the Georgia group (15) supplemented by some measurements of our own, it is possible to estimate what may be termed a 'critical velocity,' which is that velocity of linear ejection just equal to the rate of pulse wave transmission. Since both the cross sectional area of the aorta which determines linear velocity and also the pulse transmission rate are dependent upon the diastolic pressure, the critical velocity may be related to the diastolic pressure level and the ejection rate of the heart as shown in figure 10. The upper curve is that calculated for the ascending aorta; comparable calculations may be made for the descending aorta as shown below, bearing in mind, of course, that in the latter instance the total cardiac ejection must be corrected for that portion which has to be carried by the descending aorta. It will be observed that the possibilities of a water-hammer become greatest under the con-

ditions of a low diastolic pressure and a heart that has been stimulated to rapid ejection, in accordance with the situations in which water hammer pulses are most commonly observed (cf. figs. 5 and 6).

An actual illustration of this mechanism is afforded by some experiments performed by Drs. C. J. Wiggers and J. M. Werle during the course of studies on the status of cardiovascular reflexes in hemorrhagic shock, shown in figure 11. The dog had been bled to hypotensive levels some 2 hours before this experiment was performed. The pulse on the left illustrates the aortic contour before stimulation, together with the simultaneously recorded cardiometer tracing below. The downstroke in the cardiometer tracing beginning at the onset of aortic systole indicates the reduction in ventricular volume with cardiac ejection. As a test reflex, the central end of the left vagus nerve was subjected to faradic stimulation. This resulted in marked cardiac stimulation, as may be seen in the progressively steeper drop in the cardiometer recording in successive records taken at short intervals from the original recording. Accompanying this sharp drop in heart volume with ejection, there appears progressively a characteristic water-hammer spike on the anacrotic crest of the aortic pressure curve. In spite of the relatively small stroke volumes in this case, the vigorous ejection and the very low diastolic pressure result in an ejection velocity which, by calculation, is found to be significantly in excess of the theoretical critical velocity as shown in figure 10.

This concept of the anacrotic spike as a water-hammer is therefore afforded some substantial support. As would be expected, such a local phenomenon usually tends to be dampened out in transmission of the pulse to the periphery as shown in figure 5. In some instances, however, such a damping does not seem to occur and indeed the effect may be exaggerated. This would appear to be explainable as a process of regenerating the water-hammer, since, as suggested in figure 10, each successive segment of the vascular bed is potentially able to produce such a water-hammer effect if the critical velocity at any point is exceeded.

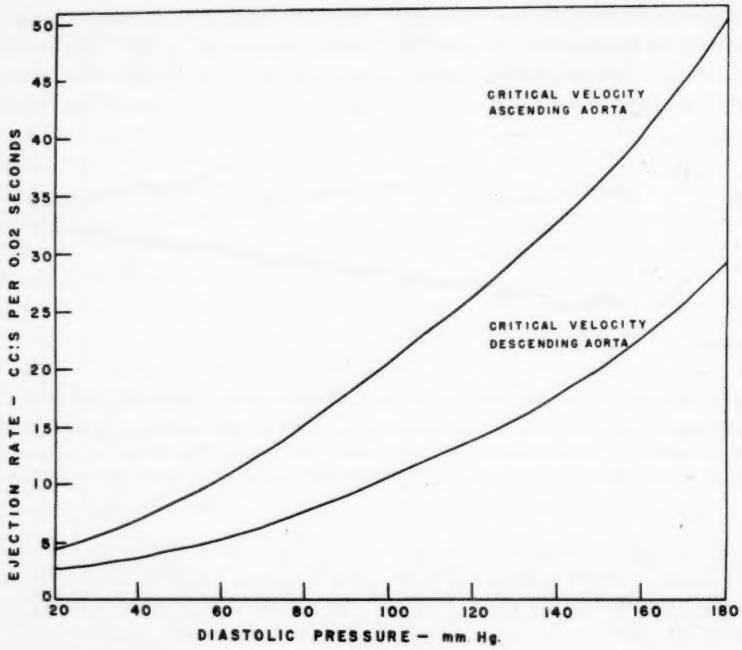


Fig. 10.

Unfortunately the curves given in figure 10 are based upon data which are not adequate for precise quantitation of this phenomenon. Substantiating data are needed of exact measurements of the pulse wave velocity in the ascending aorta, which to our knowledge are not as yet available. In addition, precise data are needed as to the cross sectional area at the mouth of the

aorta at different diastolic pressures and possibly under different mechanical and physiological states of the heart. Actually, a survey of many recordings has convinced us that the anacrotic spikes begin to appear at somewhat lower ejection rates than that predicted in figure 10 and are sometimes observed at higher pressures than expected. Nevertheless, it is felt that the present argument should be of value as a working hypothesis, emphasizing that further investigations must be carried out to determine whether this water-hammer mechanism can adequately explain the apparent discrepancy, cited earlier by Opdyke (8), between the actual course of cardiac ejection as revealed by the cardiometer and that which might be inferred from the rate of rise of pressure in the aortic Windkessel.

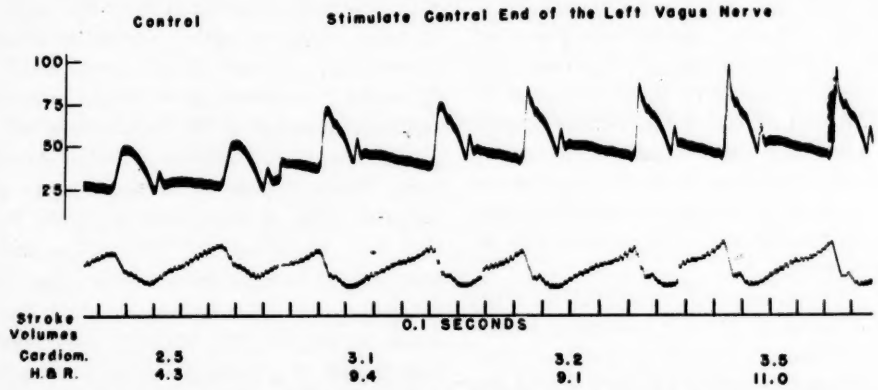


Fig. 11. RECORDINGS of aortic pressure (*above*) and cardiometer volume of both ventricles (*below*) illustrating genesis of watee hammer pulse as described in text. Figures at bottom indicate left ventricular stroke volume calculated by taking ½ volume chang indicated by cardiometer tracing, and stroke volume calculated by the method of Hamilton and Remington.

DISCUSSION

On the basis of the foregoing analysis of the factors contributing to the aortic pulse contour, it now becomes relevant to return to the question presented at the outset: Is our understanding of the factors determining pulse contour adequate to justify a rational solution of the problems of quantitation?

The initial approach must necessarily be directed towards the Windkessel concept. Opdyke (8) has already reviewed the background of attempts to define the distensibility characteristics of the aortic system. We should re-emphasize the significant advance made by the Georgia group which identified different anatomical segments that must be treated separately if accurate information regarding volume increments is to be obtained from measured pressure increments. The major unsolved problem in this connection relates to the possibility that vasomotor actions may be capable of significantly altering the distensibility characteristics of the aortic Windkessel. Much of the evidence of changes in the elastic properties of the aortic system has been based upon measured changes in the transmission rate of the foot of the pulse wave, a form of evidence which Remington and Hamilton (16, 17) have shown to be largely, if not completely, invalid. These workers have further presented some evidence (17) that the muscular component of blood vessels produces less change in the relative distensibility of the vessels than many of us have supposed. On the other hand, Wiggers and Wegria (18) attempted a direct attack on this problem by devising an instrument which would record the actual diameter of the descending thoracic aorta *in vivo*. Recordings with this instrument revealed striking changes in the distensibility characteristics of the aorta produced by reflex stimulation and by adrenalin injections. Technical difficulties encountered in this method apparently discouraged them from pursuing their studies to obtain definitive evidence as to just how these distensibility changes were produced. It therefore appears that, although the Georgia group may have some justification for minimizing the importance of vasomotor changes, a conclusive answer to this problem should be obtained before one can confidently ignore this possibility in quantitating the stroke volume from the aortic pressure pulse.

A further problem in the way of quantitation is represented by the standing wave undulations in pressure which are superimposed upon the

aortic pulse contour. Since this represents blood which is essentially bouncing around within the Windkessel, any method of quantitation which attributes these pressure changes to alterations in either cardiac ejection or peripheral drainage will correspondingly be in error. Unfortunately, though the standing wave is easy to recognize qualitatively, it appears to be difficult to define quantitatively (11). Hamilton and Remington are thoroughly cognizant of this problem and, in presenting their method for quantitation (19), express the opinion that the errors introduced by failing to correct for this factor would be small. Although in general this is true, unusual distortions of the standing wave system have been observed in shock (4). It is therefore pertinent to note that the method of quantitation presented by Hamilton and Remington appears most vulnerable to error in states of shock, as noted by Huggins (20) and others. A method for quantitating stroke volume which is valid for the normal animal but which is subject to error whenever one encounters serious disturbances in the circulation has a sharp restriction in its applicability to investigative work.

Finally, the success of quantitation must resolve the problem of the anacrotic water-hammer described above. This is emphasized in figure 11 where the actual stroke volumes are compared with those calculated by the method of Hamilton and Remington. It is obvious that the appearance of the anacrotic water-hammer seriously distorts the calculated output, as evidenced by a calculated increase in output of 156 per cent compared with an actual increase in output of only 40 per cent. One might suppose that this could be corrected by merely rounding off the curve by inspection, so as to eliminate the spike. Study of many such records, however, have convinced us that such a subjective correction of the pulse contour leads to a serious error, especially when it is remembered that the water-hammer phenomenon involves, not just a visible spike, but the entire anacrotic upstroke of the pulse.

To avoid serious errors in quantitation from pulses of this type, which are not uncommon at lower pressures and are sometimes seen at high pressures, some basis must be devised for quantitatively accounting for this phenomenon. A significant advance in this direction has been made by Peterson, and further work may well render this problem soluble.

CONCLUSIONS

On the basis of an analysis of the aortic pulse contours observed under diverse conditions, it appears that present knowledge of aortic pulse dynamics is adequate to explain in a qualitative way most if not all of the details observed. Some of the variants are readily susceptible to quantitation; others appear to demand more precise definition than is at present available. It is felt that Drs. Hamilton and Remington and their associates at Georgia deserve to be congratulated for both the significant advances which they have made in defining the variables and also their formulation of a method for calculating cardiac output which involves much greater recognition of these variants than has been true of previous methods. It is to hoped, moreover, that further resolution of some of the problems discussed here will permit a greater and more consistent degree of accuracy in quantitating the aortic pressure pulse under some of the more extreme departures from normal circulatory dynamics. It is especially in the latter situations that an accurate method of calculating the stroke volume of the heart is so sorely needed.

REFERENCES

1. Peterson, L. H., T. G. Schnable, H. F. Fitzpatrick and H. C. Bazett. *Am. J. Physiol.* 159: 585, 1949.
2. Hamilton, W. F. *Am. J. Physiol.* 141: 235, 1944.
3. Wiggers, C. J. *Physiology in Health and Disease.* Philadelphia: Lea & Febiger, 1949, 5th ed., Ch. 37, 38.
4. Alexander, R. S. and E. A. Webb. *Am. J. Physiol.* 150: 272, 1947.
5. Remington, J. W. *Federation Proc.* 11: 750, 1952.
6. Wiggers, C. J. and L. N. Katz. *Am. J. Physiol.* 58: 439, 1921.
7. Green, H. D., R. N. Lewis, N. D. Nickerson and A. L. Heller. *Am. J. Physiol.* 141: 518, 1944.
8. Opdyke, D. F. *Federation Proc.* 11: 733, 1952.
9. Lawson, H. *Am. J. Physiol.* 140: 420, 1943.
10. Hamilton, W. F. and P. Dow. *Am. J. Physiol.* 125: 48, 1939.
11. Alexander, R. S. *Am. J. Physiol.* 158: 287, 1949.
12. Alexander, R. S. In preparation.

13. PETERSON, L. H. *Federation Proc.* 11: 762, 1952.
14. MALTBY, A. B. AND C. J. WIGGERS. *Am. J. Physiol.* 100: 604, 1932.
15. REMINGTON, J. W. AND W. F. HAMILTON. *Am. J. Physiol.* 144: 546, 1945.
16. HAMILTON, W. F., J. W. REMINGTON AND P. DOW. *Am. J. Physiol.* 144: 521, 1945.
17. REMINGTON, J. W., W. F. HAMILTON AND P. DOW. *Am. J. Physiol.* 144: 536, 1945.
18. WIGGERS, C. J. AND R. WÉGRIA. *Am. J. Physiol.* 124: 603, 1938.
19. HAMILTON, W. F. AND J. W. REMINGTON. *Am. J. Physiol.* 148: 14, 1947.
20. HUGGINS, R. A. *Federation Proc.* 11: 767, 1952.

VOLUME QUANTITATION OF THE AORTIC PRESSURE PULSE[1]

JOHN W. REMINGTON

From the Department of Physiology, Medical College of Georgia, Augusta

THE arterial tree must serve two not entirely compatible functions. Designed primarily as a conduction network through which blood can be moved from the heart to all the tissues, it is composed of a great number of tubes, arranged partly in parallel and partly in series, which become progressively smaller with each branching. This arrangement compromises the second action, that of acting as a storage bellows, or 'Windkessel,' which can protect the peripheral vessels from large pressure changes. All tubes in the arterial tree have distensible walls, and hence can gain volume when subjected to pressure increase. The volume gain of each of the small arteries is, of course, quite small, but, like Lincoln's poor people, they are so numerous that their aggregate contribution to total Windkessel action is appreciable. Our estimates indicate that the extra-aortic vessels have a total volume uptake some 1.5 times that of the aorta. Hence in considering the properties of the Windkessel, we have no right to neglect any part of the arterial tree.

In figure 1 is shown a semi-diagrammatic drawing of the larger arterial vessels of a dog, in which the length values have been kept in scale. These dimensions were based upon the dissection of an animal whose arteries were injected with cement at a pressure of 100 mm. Hg (1). Before we can attempt a volume uptake prediction for such a system, this picture must be modified. First, since the vessels will gain volume only after the pulse wave has reached them, the Windkessel should be redrawn according to the time-distance of the various segments from the heart. Since the pulse wave starts slowly, and then speeds in the more distal tubes, the anatomical drawing will be foreshortened when transmission time is used as one dimension scale. Such a construction is shown at the bottom of figure 1. At a pressure of 80 mm. Hg, for instance, the pulse wave is entering the head some 40 msec. after ejection has begun, and reaches the hind foot some 90 msec. after ejection starts. Hence not only may the pulse pressure be quite different at any given instant in the various parts of the Windkessel, but it is also possible for the wave not to have arrived in the distal parts of the arterial tree before ejection is completed.

We must, then, make some sort of division of the Windkessel into two parts, viz., that which is predominately filled during cardiac systole, and that which will be filled later, during diastole. For the blood which goes to fill the latter fraction will come not from the heart but from the previously swollen proximal vessels, and this blood must not be counted twice. For this fractionation, we must have a set of reference transmission time values, varying, of course, with the pressure level.

Not many arteries of the body are so conveniently located as to allow entrance with a manometer needle, and a direct determination of transmission time. By gathering together all past records, showing pressure waves as recorded in various accessible arteries, and by making what seemed reasonable assumptions based on vessel size and length, we drew up a composite table of transmission time values through the major arterial beds. We know that dogs of different sizes will show times different from those of this table. Further, we know that the pulse wave velocity may change, even at the same pressure level, with changes in tension of the smooth muscle of the wall, and also with the heart rate (hysteresis action of the wall) (2). A true subdivision of the Windkessel would demand an assessment of transmission times for each dog studied. Since this would be too cumbersome to be practical, we merely assumed that the single average values would be applicable to all animals and to all physiological states, hazardous though such an assumption may be.

In this division of the Windkessel into systolic and diastolic fractions, the reference point was the end of cardiac ejection, arbitrarily taken to be coincident with the incisura as registered in the ascending aorta. The transmission time to the various beds will then be laid back from the incisura. Hence, in the pulses shown in figure 1, the head arteries will have developed all but the

[1] This investigation was supported by grants from the National Heart Institute, National Institutes of Health, U. S. Public Health Service, and from the Life Insurance Medical Research Fund.

last 40 msec. of the pressure pulse, while those of the hind leg will not have developed the last 90 msec. of the pressure pulse.

The next step, then, is an assessment of the volume change per unit pressure rise for the various parts of the arterial tree. Since independent distensibility measurements cannot be made in every animal, we determined such values as completely as possible for a few dogs, hoping that the average figures obtained would be applicable to other animals. First, rings taken from the aorta, and from the larger arteries such as the femoral and carotid, were subjected to stretch. The tension-length figures obtained were converted to volume-pressure terms. Second, in 4 newly killed dogs, the aortas were dissected free, the small arteries having been tied off, and the major exit arteries were closed by hemostats.

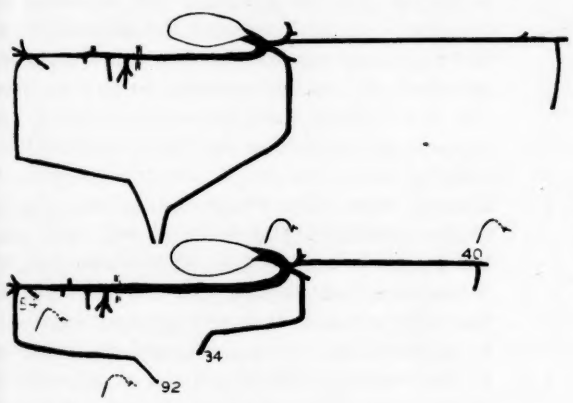


Fig. 1

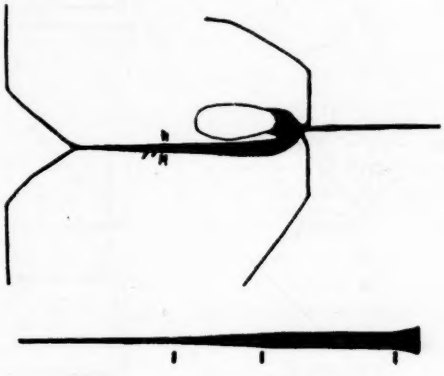


Fig. 2

Artificial injections were then made into this system, which, by opening the proper hemostats, could include the aorta alone, or the aorta plus each of the major arterial beds. The volume-pressure pulses so obtained were plotted, average values taken, and curves connecting the points smoothed in from our knowledge of the contour of the volume-pressure relation of isolated rings. It was also necessary to extrapolate these curves into pressure regions not reached by the artificial injections. The whole of an arterial bed was treated as though it consisted of a single, composite tube, with a single volume distensibility.

Much has been made, both in the past and currently, of the thesis that arterial distensibility must be conditioned largely by the initial volume. Differences in diameter may arise from two general causes. First, animals of different body sizes will obviously have arteries of different internal volume, wall thickness and length. As a working hypothesis, we assumed that the volume change in this case would be accompanied by a corresponding change in relative distensibility ($\Delta v/V\Delta p$), and that the absolute volume gain per unit pressure rise ($\Delta v/\Delta p$) would vary with the surface area of the dog. Hence stroke volumes calculated from the volume uptake table will be in terms of a square meter body surface. As will be shown later, the essential correctness of this assumption is attested by a lack of consistent deviation of the calculated stroke indexes from those given by the Fick procedure when body weights have varied from 6–24 kg.

Second, in animals of the same size, a change in vessel diameter may accompany aging, disease, or a physiological change in smooth muscle tension. The increased volume is not, in this case, accompanied by an increased number of wall fibers. Present knowledge will not allow a complete statement as to how, and to what degree, such a diameter change will affect distensibility. Our studies did indicate that when the initial volume increased, the relative distensibility decreased, with the actual volume gain remaining relatively constant (2–4). An analysis of this situation has been presented (2), and need not be repeated here. These data led to the conviction that a working method could be developed on the premise that volume uptake values were relatively independent of changes in initial diameter.

With a table for transmission times through the arterial bed, and one for volume uptake of the regions of the bed, we are now ready to diagram the arterial system as a working Windkessel. In figure 2 is shown such a picture, in

which tube length represents time-distance from the heart, and the tube diameter, the volume gain/mm. Hg pressure rise/msec. length of tube. We can made one further simplification, and say that it doesn't concern our attempted quantitation that the blood may leave the arch in four different directions, and hence fuse the whole arterial bed into one tube. As seen in the bottom figure, this tube is in the shape of a funnel, with a wide mouth in the ascending aorta, and a long stem representing the abdominal aorta and the arteries of the viscera and legs. The marks below the funnel show the limits of the four great regions of the arterial bed, i.e. the arch; the thoracic, head and fore-leg vessels; the abdominal aorta and visceral arteries; and the hind-leg vessels.

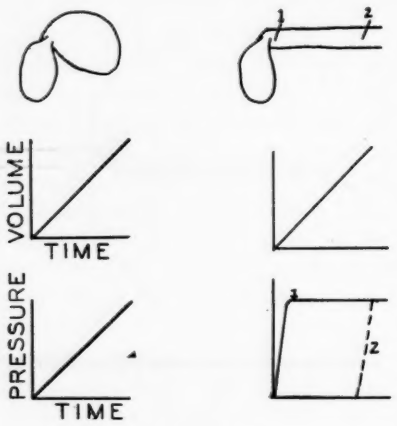


Fig. 3

This funnel shape may at first sight appear inconsistent with the common picture of an ever increasing arterial volume as the arterial vessels branch. Yet the peripheral tubes are relatively stiff, with small volume uptake values per unit length of tube. While the Windkessel function of all of them may be appreciable, when uptake is expressed in terms of time-length of tube, their distensibility is less than that of the short, compact arch.

To illustrate the effect of the Windkessel architecture upon a pressure pulse, let us assume a cardiac pump which could attain instantaneously a constant ejection velocity. Let us also suppose that ejection were into a simple Windkessel whose walls were constructed of an elastic material similar to an aortic wall (but not to rubber) such that the volume-pressure relation was constant. Since pressure would be dissipated equally in all directions, with a linear ejection curve the pressure should also rise in a linear manner, as diagrammed in figure 3. Next, let us suppose that the ejection were into a tube of uniform diameter and uniform wall distensibility. As ejection began, blood would be forced into the first segment of tube. A small part of the energy of this blood would be spent in accelerating the fluid column. The rest would stretch the distensible walls and construct the front of a pressure pulse. The recoil force of the walls would maintain the accelerated rate of blood movement down the tube. The front of this moving fluid wave, pushing on the more slowly moving blood ahead, would in turn stretch the walls of the next section of tube. Assuming no frictional energy loss, just the same amount of blood is required to establish the pressure rise in the second segment of tube as in the first. The first segment, then, is yielding to the more proximal segment the same volume that it is receiving from the heart, and its pressure level will remain constant. So long as ejection is continued constant, and so long as the wave is still traversing the tube, this constant pressure must continue. If the tube is open at the end, then when the pressure pulse arrives at the terminal segment, blood will spill out, leaving behind a negative wave which will be propagated back toward the first segment. If the tube is closed, then the pressure wave will be reflected back upon itself, and the pressure in the terminal segment will rise to twice what it had been. This positive reflected wave will now move back through the tube. Until it arrives, the pressure in the first segment must remain unchanged, and this segment should have no fore-knowledge of the length of the tube.

In the arterial funnel, less blood is required in the narrows to establish the original pressure front than was needed in the mouth. It would be reasonable to suppose that multiple reflections would be set up from the walls of the narrows, and that the ascending aorta would receive many returning waves which would raise its pressure. Yet small ripplets are not to be seen on a well recorded natural pulse. On less sound theoretical foundation, we could simply say that blood not required to establish the pressure front in the narrows would be left behind, and when added to that received from the heart, would cause a further pressure increase. In either case, the pressure will not remain constant after the initial steep rise, as was true in the uniform tube, but will continue to rise gradually.

Three Windkessel factors will influence the shape of the pressure curve. First, the rate of the initial upstroke will reflect the distensibility of the first time segment in the mouth of the funnel. Arterial distensibilities will change as the pressure is altered. With a pressure rise, the aortic wall becomes stiffer, so that a given pulse pressure will require less volume gain. At the same time, the pulse wave velocity is increased, so that a greater length of tube is being pervaded by the pulse wave in a given time interval. The actual change in distensibility per time-length of tube will depend upon the relative rates of change of these opposing factors, which is different at different pressure ranges, and in different parts of the arterial tree. Usually, the transmission time is decreased more than is the absolute distensibility, so that the net change is for a greater volume uptake for a given time-segment. There is, however, little change in the very mouth of the funnel. As indicated in figure 4, there is little difference in initial pressure upstroke at pressure values of 80 and 120 mm. Hg, and that at 20 mm. Hg is but slightly steeper.

The second factor concerns the funnel arrangement proper. When the front of the pressure wave enters the narrows, blood being 'dammed back' in the upper funnel raises its pressure. The amount of this pressure rise will reflect the amount of narrowing, and the time necessary for the wave to enter, and to pass through, the tapered portion. The form of the funnel is changed at different pressure levels. The upward 'bowing' in pressure becomes more pronounced as pressure is lowered. Since the duration of the 'bow' is also increased, the actual pressure elevation at any given time instant may not be larger. Since the initial pressure upstroke merges into the 'bow,' the composite effect is for a more gradual pressure ascent in early systole at high pressures than at low.

Lastly, when the wave front has traversed the Windkessel, pressure differences between the ends of the funnel are greatly reduced. On the thesis that we will not be concerned with the returning reflected wave, we are now working with a filled Windkessel which will approach, in its pressure response, the previously described simple Windkessel.

Working on the assumption that the pressure front established in the ascending aorta will be propagated inviolate through the arterial bed, and by actually recording the pressure changes in the mouth of the funnel, we should be able by proper book-keeping, to construct a cardiac ejection curve corresponding to any given pressure pulse. We know, however, that the pressure curve is changed in its progress through the aorta. A pulse in the femoral artery, for instance, shows a pressure rise which continues steep for a longer interval, and a higher pulse pressure. If we assume that this change denotes a fusion between the primary wave and the reflected wave, this change in form need not concern us, for the volume involved in the reflected wave has been taken into account, earlier in systole.

The validity of the contour of such a derived ejection curve has not been directly proven. In general, such curves have a form quite in keeping with published cardiometer curves. If we attempt the reverse process, and derive a pressure pulse from an arbitrarily selected cardiometer curve,

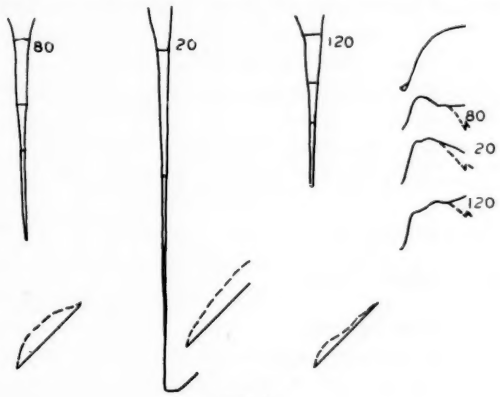


Fig. 4

we see at once that the pulse will differ from a natural one in a few, but perhaps very significant, respects. Such derivations are shown at the right side of figure 4. As predicted, the initial pressure upstroke is not greatly different at 80 and 120 mm. Hg, but is more rapid at 20 mm. Hg. The pressure 'bowing' is more pronounced at the lower pressure values. Only at 120 mm. Hg does the funnel become filled during systole, so that the pressure rise assumes a linear rate of increase. The slope of this increase is less than that which would have been effected earlier, for the ejection rate has slowed quite appreciably by this time.

When the pressure pulse reaches the arterioles, the drainage rate is accelerated, which robs the Windkessel of some of its stored blood. Since an appreciable time interval is required for the pulse to reach the bulk of the arterioles, the drainage factor will affect only the last part of the pressure pulse. In the living animal, a high pressure is

usually a reflection of vasoconstriction, with a reduced drainage rate. Hence the Windkessel loss is least, and the pressure fall because of drainage (broken line of fig. 4) is not large. Conversely, at low pressure levels the drainage rate is larger, and the pre-incisural pressure fall may be quite extreme.

The natural pulse shows a very steep initial upstroke of constant slope, lasting some 10–30 msec., which is followed by an inflection and a mergence into a more gradual pressure rise to the systolic peak. This pattern is seen on none of these derived pulses. We are confronted here with a serious problem of interpretation of the cardiometer curve, for at the beginning of ejection the cardiometer record shows a negative dip presumably reflecting a movement of the heart into the recording chamber. It is usually assumed that the actual volume curve will start slowly, and will fuse with the cardiometer curve at the end of this dip. If, however, the rapid initial pressure rise is associated with the expected volume increase, then we must conclude that the initial ejection is much more rapid than this assumption would allow, and that the cardiometer curve is actually showing an appreciable lag behind the ejection curve. A more rapid initial ejection might perhaps be expected, for cardiac contraction is well under way when the valves open, and the valves open suddenly. However, we must admit that the initial pressure rise may also be reflecting a hysteresis of the aortic wall, with a reluctance to stretch (2). This would produce a pressure increase with relatively small volume gain. After this, when the wall starts to give more readily, ejected blood must make up this prior deficit, and the rate of pressure rise will be less. The greatest hysteresis effect is noted at low pressure values, the least at high pressures, which is in keeping with the steeper initial segment with low pressure pulses.

The effect of the Windkessel form upon the pulse contour can be illustrated with some natural pulses. In figure 5, top, is shown a pulse before and after the intravenous injection of epinephrine. The figures given over the pulses are the pressure values, the length of systole in msec., and the length of the cycle. As the pressure increases, the anacrotic slope becomes more gradual. This is in part attributable to the 'shortened' Windkessel and to the lessened 'bow' due to funnel structure. It might also be due in part to a lessened hysteresis effect seen at high pressures. The duration of the 'bow' is less, and the pressure rises in almost linear manner in late systole. Further, when the heart is faced with an elevated resistance to ejection, not only is the stroke volume reduced but the contraction pattern is altered. Instead of an initial burst of activity, the ventricle ejects more slowly but at a more steady rate until late in systole (5). How much of this change is intrinsic in the heart and how much marks the influence of the Windkessel upon the ejection curve, and indirectly upon cardiac activity, cannot be answered.

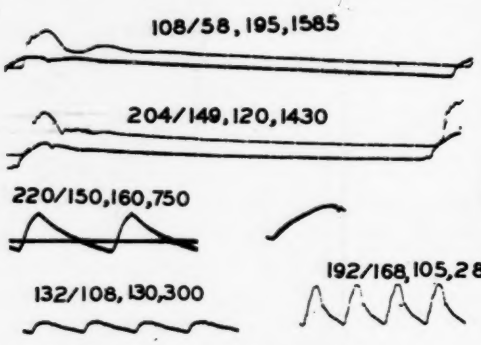


Fig. 5

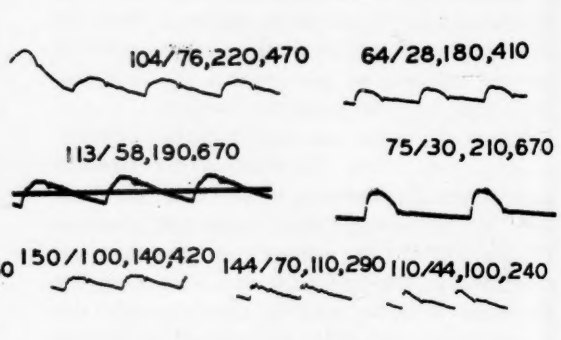


Fig. 6

An even greater degree of anacrotic slowing and a better sustained pressure increase is seen in the third pulse, taken by needle from the carotid artery of a dog with renal hypertension. The camera speed was considerably less than for the top records. A drawing of the systolic portion of the pulse at the faster camera speed is therefore given at the right side of the actual record.

Of course, the pressure cannot be supported in late systole unless cardiac ejection is continued at a moderately good rate. In all of the above cases the heart rate was quite slow, the ventricular volume large (6), and the ejection was

sustained. The curves at the bottom of figure 5 are from an atropinized animal before and after an epinephrine injection. Here the heart rate was rapid, the heart volume small, and ejection very much slowed toward the end of systole. The pulse therefore tends to collapse before the incisura.

In figure 6 are shown some low pressure pulses. The first shows the result of releasing an aortic occlusion. During the occlusion period the pressure was high, the heart rate slow, and the amount of residual cardiac blood large. When the resistance to ejection was suddenly reduced, ejection could be well sustained throughout systole (7). Later, as the heart became smaller (second series of pulses, first line) the pre-incisural pressure fall was more extreme. As might be expected, with the 'lengthened' Windkessel, the initial pressure rise was steeper. The pulses of the second row, taken before and after the injection of diodrast, which dilates the arterioles, illustrates these changes even more strikingly.

If, as the arterial pressure falls, the heart is at the same time stimulated, these changes in pulse form become exaggerated. The last series of pulses show the effects of an intravenous injection of iso-propyl-arterenol, which shortens ejection and accentuates the initial ejection spurt (5). The anacrotic rise is very steep, and is usually followed by aortic vibrations. Ejection fails early, so that the pre-incisural pressure fall is marked. This fall is made even larger by the increased rate of systolic drainage.

Since our procedure for calculating an ejection curve, and a stroke index, from the pressure pulse rests upon a number of assumptions, our concepts will be valid only if these stroke indexes show quantitative agreement with those given by an acceptable reference method of measuring cardiac output. To make a valid comparison is, however, a most difficult problem. The reference methods are measuring mean flow through the pulmonary circuit. The time interval over which this flow is being determined cannot be specifically fixed. The pulse contour method presumably is recording left heart output, beat by beat. It would indicate that in any animal where the heart rate and pulse pressures are not constant, and such is true of almost every dog in good condition, a so-called steady flow is an imaginary thing. By averaging a great number of pulses, we might hope to arrive at a mean flow which might be taken to be the same as that being recorded by the Fick procedure, or by the dye injection technique. The uncertainty of what beats to measure, and how many of them should go into this average, does leave the investigator with a feeling of perplexity. In many cases, the stroke volume may change very appreciably during the time course of the blood withdrawal for the Fick or dye determination. We can only hope that by making a large number of flow comparisons, some index can be had to the validity of the contour method through averages. A great many such comparisons have now been presented, from several different laboratories. In most respects, the results have been quite similar. Under a rather wide range of experimental conditions, and in open as well as closed chest dogs, the average error of the pulse contour stroke index as against another flow measure is a little over ± 10 per cent. The range of variation is of the order of ± 40 per cent.

Three questions must be answered by these comparison experiments. First, does an individual animal show a consistent difference, indicating a consistent variation from the two basic tables because of body size or inbred differences in vessel size and distensibility? In our experience, such consistent differences are quite rare. In any given dog, differences between the flow indicated by the pulse contour calculation and that given by the Fick have appeared to be random.

Second, are there systematic deviations at certain pressure ranges, which would indicate that the volume uptake table had an intrinsic error in its smoothed curves? We noted several years ago a small systematic error at pressure ranges of 70–100 mm. Hg, with the pulse contour value being about 7 per cent higher than that given by the Fick. Longino and Gregg (8) described a more serious error at high pressure values, in the opposite direction. Our average pulse contour values at high pressures also tend to be somewhat low, but rather than dealing with a truly systematic error, we seem to have a bimodal series, with one group of pulses fitting the measured flows, and another group showing the low pulse contour stroke volumes. This is illustrated by the results shown in *column H*, table 2. We have not been able to sort out the low pulse contour stroke volumes on the basis of dog size, blood pressure level, heart rate or resistance. There does seem to be a trend depending upon the manner in which the drug was given. A dog given an infusion, with maintained high pressure values, or a dog hypertensive because of buffer nerve section, tends to show better agreement with the Fick results than one given

a single intravenous injection of a rapidly acting drug, such as epinephrine. It is probable *a priori*, although certainly not proven, that, in our series, the low pulse contour stroke volumes are reflecting an inability of the Fick procedure to follow rapid flow changes.

On the basis of our survey, we would make some minor adjustments of the volume uptake table to render the empirical fit better with the

TABLE 1. NET VOLUME UPTAKE OF ARTERIAL TREE AT VARIOUS PRESSURES

PRESSURE	ARCH	HEAD AND THORACIC	VISCERA AND ABDOMINAL	LEGS
mm. Hg	*cc.*	*cc.*	*cc.*	*cc.*
20	1.4	3.5	1.3	1.1
40	4.6	9.9	3.7	3.4
60	8.1	17.6	6.5	6.1
80	11.5	25.5	8.9	8.3
100	14.6	32.9	10.9	10.1
120	17.6	39.3	12.6	12.0
140	19.9	44.5	14.1	13.2
160	21.6	48.0	15.3	14.3
180	22.8	50.6	16.4	15.2
200	23.7	52.4	17.4	16.0
220	24.6	54.0	18.3	16 6
240	25.4	55.5	19.0	17.2

That they are present would indicate that cardiac index calculated from pulse contours should not be expressed in more than two figures. It would also seem that our calculation of systolic drainage need not be so meticulous as first proposed. The reading of a great many pulses has revealed that the mean pressure level during effective systole ($Ts - Tw$) is seldom more than a mm. Hg or so different from the mean level during effective diastole ($Td + Tw$). With low pressure pulses, where, for example, the pulse pressure being high and the diastolic slope low, we might expect the greatest difference between these means, Tw is also long, so that effective systole occupies only the first part of total systole. If we assume that a difference between these mean pressures is not practically critical, then systolic drainage/diastolic drainage will simply equal $(Ts - Tw)/(Td + Tw)$, or equal $(Ts - Tw)/(Tc - Ts + Tw)$, where Tc is the cycle length. Since diastolic drainage is taken to equal systolic uptake, U, then

$$\text{Systolic drainage} = U\,(Ts - Tw)/(Tc - Ts + Tw), \text{ and}$$

$$\text{Stroke Index} = U + \text{Drainage} = UTc/(Tc - Ts + Tw).$$

TABLE 2. DISTRIBUTION OF 'GOOD' AND 'BAD' PULSE CONTOURS UNDER VARIOUS EXPERIMENTAL CONDITIONS

Contour/Fick × 100	A CONTROLS		B ACUTE HEMORRHAGE		C ADRENAL CRISIS		D VENOUS OCCLUSION		E INTRAP. GLUCOSE		F SALINE INFUSION		G VASOCONSTRICTORS		H VASODILATORS		TOTAL A–E		TOTAL F–H	
	Good	Bad	Good	Bad	Good	Bad	Good	Bad	Good	Bad	Good	Bad	Good	Bad	Good	Bad	Good	Bad	Good	Bad
60–79	11	0	4	1	3	2	4	0	10	0	3	0	21	0	2	2	32	3	26	2
80–99	24	0	26	4	5	1	4	3	8	4	1	1	14	0	13	4	67	12	28	5
100–119	32	0	18	1	7	0	13	1	26	2	4	0	20	0	8	1	96	4	32	1
120–139	5	0	7	1	8	0	8	2	5	6	2	0	6	0	8	2	33	9	16	2
Over 140	1	0	1	3	2	2	4	17	7	23	0	2	3	0	5	21	15	45	8	23
Total.....	73	0	56	10	25	5	33	23	56	35	10	3	64	0	36	30	243	73	110	33
Av. error, %.....	13.1		12.7	35.5	21.2	63.0	24.4	73.2	22.1	69.1	19.5	85.0	20.7		21.2	75.2	17.5	65.4	20.8	76.1

Fick flow. Such a revised set of data are given in table 1.

The reasons for the random scatter of pulse contour values away from the Fick determinations that are observed during the course of various experiments cannot be specifically delineated. If the fault is with the pulse contour method, we can only point out again the various assumptions upon which the method is based. If flow be expressed in terms of cc/sec/m.^{2},

$$F = U/(Tc - Ts + Tw)$$

The third, and really most important question, is whether the pulse contour method can really be safely relied upon to follow stroke volume from beat to beat. So long as no other direct reference method can register single stroke volumes this question cannot be settled. With but rare ex-

ception, however, no matter how close the actual quantitative value given by the pulse contour method is to that of the Fick, directional changes have tended to be parallel. This is just as true when the pulses have developed the 'distortion factor' (to be discussed later) as when, in the usual experimental conditions, they are seemingly accurate. We do not believe that in any of the experiments in which we have used the method, and these cover quite a wide range of experimental conditions, we have been deceived as to the directional pattern of change in either flow or resistance. Further, while we admit freely the amount of scatter which is at times seen, we do not believe that this scatter invalidates the usefulness of the method in following beat-to-beat changes in stroke index. Much useful physiological information has been made available by the technique. In almost all respects, this information complements evidence presented in the past by different techniques. We will, therefore, continue to use the method to measure cardiac output for most types of experiments.

It is clear, however, that there are some unique experimental conditions where the pulse contour value becomes suddenly, and will usually remain, very much greater than the actual flow. Some glimpses into the situation were first observed in our study of circulatory failure in adrenalectomized dogs (9). This discrepancy was noted, but dismissed as another peculiarity of this most perverse of experimental creatures. Two years ago, Dr. Huggins called our attention to a very large systematic error which could be almost universally produced by the insertion of a pump, or a large rotameter, into the thoracic cavas, a finding highly suggestive of the open-chest results of Duomarco and associates (10). With disbelief, I went to Texas for a summer to see where his technique had gone wrong. A series of animals done together certainly convinced me that it is quite possible for a failing open-chest dog to show false pulse contour stroke volumes and rather bizarre looking pulse forms. Our immediate problem, upon my return to Georgia, was to see if a comparable situation could be produced in a closed-chest dog, and if we could define in part the conditions under which the 'distorted' pulse would develop.

In our first approach we tried, as did Huggins and Smith in their open-chest work, a partial occlusion of the inferior cava. A long-handled adjustable brass clamp was placed around the inferior cava, the handle of which was brought out through the chest wall. The thorax was then closed and evacuated and the dog allowed to resume its own respiration. A complete caval occlusion produced a prompt and severe fall in pressure, a cardio-acceleration and a severe reduction in flow by both pulse contour and Fick methods. If sustained, the animal failed rather shortly of cardiac insufficiency. A partial occlusion produced much less pressure fall, although full cardio-acceleration occurred. In about half the cases, the cardiac index by the pulse contour method fell only slightly, while that given by the Fick procedure dropped to extremely low levels. In the other cases, both flows fell to the same degree. We were very unlucky in predicting just when, and in what animals, the disagreement between the two methods would develop.

Fortunately, we found by accident that when fatal shock developed after the intraperitoneal injection of a large amount of isotonic glucose solution, the cardiac output by pulse contour calculation was usually appreciably higher than that given by the Fick procedure. The flows by the latter method were extremely low—lower than those seen in dogs dying after hemorrhage or traumatic procedures. The pulse contour calculation might indicate a normal cardiac index, however. This glucose injection technique therefore proved a useful tool by which discrepant pulses could be produced. We found, for example, that pretreatment with Dibenamine could not prevent the development of the discrepancy, nor could vagotomy or atropinization. In some cases, the discrepancy could be corrected by the intravenous injection of hypertonic salt solution. The animals in which this could be accomplished were those that survived. When the discrepancy could not be eliminated, neither was the fatal shock reversed. The development of the discrepancy was never a gradual phenomenon; in fact, we do not have a record actually showing the change. At one sampling period it was not present, and at the next, it was. There had been a change in the pulse form in the interim. Usually discrepant pulses could be recognized by the form of the contour.

At this stage we were faced with two general problems. First, did the appearance of discrepant contours mean that the pulse contour method could no longer be trusted? We could, of course, rule out the types of experiments in which discrepant results might appear. It would be better,

however, to establish some objective criterion by which pulses, divorced from their experimental context, could be sorted into those whose quantitation should be acceptable from those which should be discarded. The second question concerned the description of what could have changed in arterial dynamics to allow the development of a pulse contour with pressure values greatly in excess of those expected from the small stroke volume. The failure of the pulse contour method in these conditions has certainly offered a clear challenge to reinvestigate all of our basic

Fig. 7

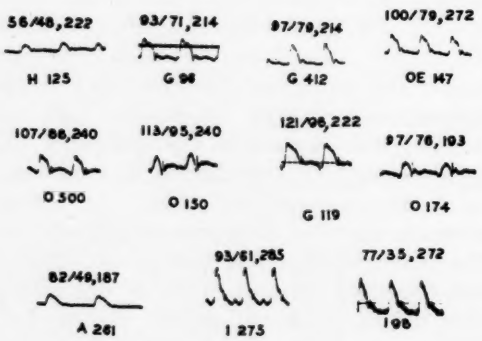


Fig. 8

premises as to how pulses are formed, and how the aorta works. This challenge has stimulated our work of the past two years.

To attack the first problem, a card file was made of some 300 pulses, each with its Fick comparison stroke index, on which cards were recorded all objective measurements that seemed likely to bear on the situation. These cards were repeatedly sorted, and many rules for separating 'good' from 'bad' contours were tested. None of them could give a universal sorting. The best rule we could find was a rather simple one. If a ruler is laid between the end diastolic and the mean incisural pressure levels, and a line extrapolated back to the anacrotic rise, an intercept over one half the pulse pressure denotes a 'good' pulse (fig. 7). If the intercept is less, but if the diastolic pressure is 30 mm. Hg or below, where relatively flat diastolic slopes might be expected, the pulse is still labeled good. All other pulses are labeled 'bad.' Some discrepant pulses did show diastolic pressure levels below 30 mm. Hg, but this modification of the general rule also 'saved' many post-hemorrhage pulses whose quantitation was sound, and whose pulse forms did not show the other characteristics of the distorted pulses. These include a very rapid initial pressure rise, a pressure peak in early systole, and a severe pre-incisural fall in pressure. Some illustrations of 'bad' pulses are shown in figure 8. Above each pulse are given its pressure value and the heart rate. Below each pulse is a letter to depict the experiment involved (H for hemorrhage, O for caval occlusion, A for adrenal crisis, I for intravenous iso-propyl-arterenol, and G for intraperitoneal glucose injections), and a number giving the ratio of pulse contour flow/Fick flow, times 100. The post-hemorrhage contour was labeled 'bad' because of the relatively high pressure values, but it did not show the other features of the distorted pulse, and its agreement with the Fick was good.

It should be stressed that the results presented here were obtained in our deliberate search for techniques which would cause the development of distorted contours. They show, therefore, poorer agreement with the Fick procedure than that obtained in our past series (11, 12), or than would be found, we believe, in future series where the method is used to describe flow changes in dogs in good condition.

The results of the pulse sorting are given in table 2. In the control group are all animals in the control period, or between drug injections, etc. In *column B* are shown pulses taken during the course of a step-wise hemorrhage, continued until death. In these two groups, agreement between pulse contour and Fick methods was comparable to that in past series. Three types of fatal shock experiments are given in *columns C* to *E*. The animals were in a relatively stable state, and the severe disagreement between flow results, when present, persisted over several repeated determinations. Of the 163 'good' pulses, 68 per cent were within ±20 per cent of the Fick value, and 94 per cent were within ±40 per cent.

Of the 'bad' pulses, 38 per cent were within ±40 per cent, and the rest all had inordinately high pulse contour stroke volumes.

In *column F* are given pulses taken from an experiment where 5–8 l. 0.9% NaCl solution was infused intravenously over the course of 2–3 hours. The infusion was stopped during the Fick determination, but flow conditions could hardly have been called stable. Of 3 dogs used, all died, presumably of anemia, and 1 of the 3 developed distorted contours. *Column G* represents results obtained after the injection of vasoconstrictor drugs, either by infusion or by a single injection. In no case did any bad contours develop. In *column H* are shown the results obtained with vasodilators, which included isopropyl-arterenol, histamine, acetylcholine and adenylic acid. In many cases where the heart was markedly stimulated, the pulse contour had all the features of the distorted pulse seen in shock states. The calculated flow from such contours was very high, while the Fick method showed little change from the pre-injection level. If a Fick determination under such acute changes be taken as valid, this marks the first case where distorted contours developed for a short period of time and then, as the drug action wore off, disappeared spontaneously.

In *experiments F* to *H*, where flow changes were acute, 93 per cent of the 'good' pulses were within ±40 per cent of the Fick value. Of the 'bad' pulses, 30 per cent were within this range. It would seem, then, that the simple rule for sorting pulses had, in the main, been effective.

Despite the fact that the diastolic pressure slope is relatively flat, the pressure values of distorted pulses are moderately high. As shown in table 3, the mean diastolic pressure for all pulses showing stroke volumes more than 40 per cent higher than the Fick value was 72 mm. Hg. They all reveal very fast heart rates, and also very brief durations of systole. Both are signs of marked cardiac stimulation (5).

When we try to offer explanations for this type of pulse we enter into the field of speculation. The immediate suggestion is that the whole, or some part, of the arterial tree has become less distensible. A spasm of outlying arteries, with the virtual elimination of both uptake and drainage, might produce such an effect. It is difficult to rationalize such a hypothesis with the fact that the distorted pulse can seemingly be developed by dilator drugs, and that a sympathetic blocking agent cannot prevent its development. There is further evidence that the distensibility change

TABLE 3. DISTRIBUTION OF CONTOUR/FICK VALUES FOR DISTORTED PULSE CONTOURS ACCORDING TO DIASTOLIC PRESSURE, HEART RATE AND SYSTOLE DURATION

DIASTOLIC PRESSURE	NO. PULSES		HEART RATE/MIN.	NO. PULSES		SYSTOLE DURATION	NO. PULSES	
	±40%	>40%		±40%	>40%		±40%	>40%
mm. Hg.						*msec.*		
30–39	6	13	160–169	4	0	Over 130	4	0
40–49	4	15	170–179	1	4	120–129	4	3
50–59	6	16	180–189	3	3	110–119	8	7
60–69	4	9	190–199	1	3	100–109	8	7
70–79	6	10	200–209	10	8	90–99	2	14
80–89	4	6	210–219	1	12	80–89	5	26
90–99	4	2	220–229	11	15	70–79	6	10
100–109	4	0	230–239	1	11	60–69	1	5
110–119	0	1	Over 240	6	16			

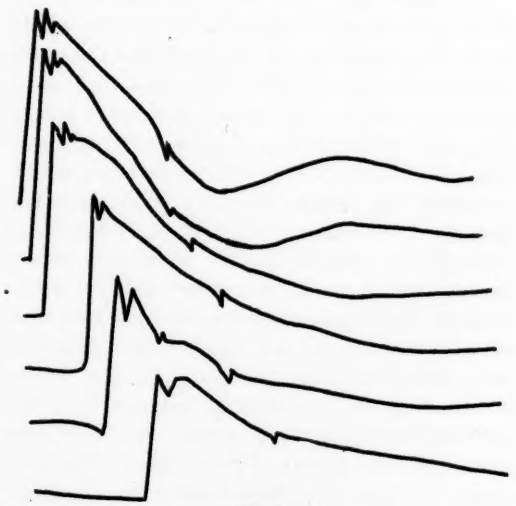


Fig. 9

must be within the aorta itself. When, for example, the progress of the distorted pulse following iso-propyl-arterenol injection is followed through the aorta, as shown in figure 9, there is no evidence for a localized change in wave form. Illustrated are, from top to bottom, pulses taken from the ascending aorta, crown of the arch, the descending part of the arch, the mid-thoracic aorta, the diaphragm region, and the abdominal aorta. The steep anacrotic limb, and the general pulse form are propagated with but minor change. There is a small but steady reduction in pulse pressure. The peripheral pulse shows no reflected waves, but the central pulse may or may not show a rather well developed secondary wave. The flat diastole of the central pulse is lost in the thoracic aorta, so that all pulses peripheral to this do show diastolic drainage slopes.

There is no direct challenge here to the premise upon which the pulse contour method is founded, that a pressure pulse will be propagated inviolate. This distorted pulse really fits this premise better than would a normal one, which does show wave reflections and a heightening of pulse pressure in the peripheral vessels.

Has there been a tone change which could produce a changed distensibility for the whole aorta? The suddenness of appearance of the pulse distortion would favor the inference that some reflex had initiated the change. Yet nerve blocks have not prevented the change, nor have any suggestions arisen as to where this reflex might originate. Our studies on aortic rings indicated that a muscle contraction would, by taking tension off the less distensible connective tissue fibers, actually make the vessel wall more extensible. In the present series, we have seen that vasoconstrictors did not produce the distorted pulse form, but that vasodilators might. There is, therefore, nothing here to indicate that our previous conception was wrong, and that if a tone change is present, it is a dilation. Measurements of pulse wave transmission time in dogs bearing rotameters, and in dogs in shock following intraperitoneal glucose injections, indicate no significant change as pulse distortion appears. If there has been a tone change, pulse wave velocity has not indicated it. The problem of what governs pulse wave velocity is of course far from settled, and certainly much more is involved here than the text-books, quoting Bramwell and Hill, would have us believe.

That the flat diastoles of the distorted pulse are lost in the thoracic aorta indicates that the contour of the central pulse might be conditioned by the appearance of a negative wave. While it must be admitted that our concepts of the genesis of the standing wave system are at present quite unsettled, it should be remembered that intrinsically the pulse contour method is concerned with the systolic portion of the curve, not the diastolic. Our problem is to explain why, in these conditions where the distorted pulse develops, the systolic pressure is higher than it should be for the actual stroke volume. Unless some new insight into wave mechanics can explain this fact, the problem of wave transformation is not relevant to the present discussion.

A steep anacrotic rise in pressure, particularly when followed by aortic vibrations, does suggest a rapid initial ejection spurt and a relatively stiff aorta. With the latter we are approaching the conditions of a rigid pipe system, where a sudden acceleration of blood may cause a large pressure change. That such a 'water-hammer' effect could support an abnormally high pulse pressure for more than half the duration of systole would seem rather unbelievable. If we could attribute the relative stiffness to the hysteresis factor, as originally described, then we would also have to conclude that the reluctance to stretch could now persist for a relatively long interval. What could produce such a change is conjectural.

In summary, the pulse contour method of calculating stroke volume has a demonstrated practical significance in its ability to describe stroke volume changes under most experimental conditions. Its accuracy is considerably better than might be anticipated from the assumptions upon which it is based. The method appears reliable enough to allow its use under the great majority of conditions, and for the purpose for which it is uniquely fitted, i.e. to measure left heart stroke volume beat by beat. There are certain peculiar conditions, however, where the method fails badly. In the main, these situations can be recognized from the contour of the pressure pulse itself. The genesis of the distorted pulse contour, whose calculated stroke volume is not to be trusted, is not understood. Since in final essence, the method is empirical, its application to quite different situations than those in which quantitative accuracy has been established is not judicious.

REFERENCES

1. REMINGTON, J. W. AND W. F. HAMILTON. *Am. J. Physiol.* 148: 25, 1947.
2. REMINGTON, J. W., W. F. HAMILTON AND P. DOW. *Am. J. Physiol.* 144: 536, 1945.
3. HAMILTON, W. F., J. W. REMINGTON AND P. DOW. *Am. J. Physiol.* 144: 521, 1945.
4. REMINGTON, J. W., C. R. NOBACK, W. F. HAMILTON AND J. J. GOLD. *Am. J. Physiol.* 153: 298, 1948.
5. REMINGTON, J. W., W. F. HAMILTON AND R. P. AHLQUIST. *Am. J. Physiol.* 154: 6, 1948.
6. HAMILTON, W. F., J. W. REMINGTON AND W. F. HAMILTON, JR. *Am. J. Physiol.* 163: 260, 1950.
7. HAMILTON, W. F. AND J. W. REMINGTON. *Am. J. Physiol.* 153: 287, 1948.
8. LONGINO, F. H. AND D. E. GREGG. *Am. J. Physiol.* 167: 721, 1951.
9. REMINGTON, J. W. *Am. J. Physiol.* 165: 306, 1951.
10. DUOMARCO, J. L., W. H. DILLON AND C. J. WIGGERS. *Am. J. Physiol.* 154: 290, 1948.
11. HAMILTON, W. F. AND J. W. REMINGTON. *Am. J. Physiol.* 148: 14, 1947.
12. REMINGTON, J. W., W. F. HAMILTON, N. C. WHEELER AND W. F. HAMILTON, JR. *Am. J. Physiol.* 159: 379, 1949.

CERTAIN PHYSICAL CHARACTERISTICS OF THE CARDIOVASCULAR SYSTEM AND THEIR SIGNIFICANCE IN THE PROBLEM OF CALCULATING STROKE VOLUME FROM THE ARTERIAL PULSE[1]

LYSLE H. PETERSON

From the Department of Physiology, University of Pennsylvania School of Medicine, Philadelphia

THE behavior of the vascular system remains one of the most complex problems of fluid dynamics. The conduits comprising the circuit vary in character and arrangement from the root of the aorta to the right atrium. The processes of aging cause changes which, from individual to individual, are irregular and variable. This continuous and irregular variation of anatomy and histology influences wall thickness, vessel diameter and the reaction to dynamic stress. These facts require that the circuit be treated analytically as a distributed system.[2]

A well established principle is utilized by engineers confronted with the problem of analyzing the behavior of a distributed, perhaps alinear, system. The engineer devises a method to drive the system with a force possessing certain defined characterizations; then analyzes the system's response to this driving force. Only after such studies are made can it be decided whether or not the system may be considered as a lumped arrangement and thus whether it can be treated simply, i.e., with a single degree of freedom. The concept of the 'Windkessel,' applied to the arterial circuit, is such a simplification. To my knowledge, adequate studies have not been done to justify such an assumption.

A device, herein briefly described, has been constructed to drive the vascular system. The method allows for a separate study of the effects of fluid moving into the vessel under two conditions, acceleration and constant velocity. The application of these separated forces of acceleration and constant velocity is linear. These properties have been essential and have pointed out to us certain highly significant characteristics of hemodynamics.

The method was applied first to the study of the reflex, mechanical and humoral influences on the 'distensibility' of veins (venomotor activity) (1). These vessels received first attention primarily because of impatient curiosity. Although this area of the cardiovascular circuit is the most distensible, controls the dynamic fate of a large proportion of the circulating blood volume, and supplies the *diastolic* load of the heart, almost nothing is known of the dynamics of these vessels. This activity must be well integrated with the remainder of the circulation. More recently the same method, with certain modifications, has been applied to the arterial system. In retrospect this order of study was most fortunate, for without the information gained from the venous vessels the interpretation of data obtained from the arteries would have been more difficult. The method also provides a means of calculating stroke volume.

The essential purpose of the device is to inject into the vessel a volume of saline solution. The time relationships of this volume pulse are most important and have been made to have the following characteristics: *1*) The velocity of saline entering the vessel changes from zero to its maximum in as short a time as possible, i.e., acceleration; *2*) having reached that maximum, the velocity then remains constant for a definite and reproducible time period; *3*) at the end of this period, the velocity changes from its maximum value to zero in as short a period of time as possible, i.e., deceleration; *4*) the pulse provides enough energy and volume to adequately drive the system.

These characteristics describe, ideally, a velocity square wave and a right triangle of volume change as noted in figure 1. Such a device has been constructed and its performance satisfactorily approaches the ideal behavior. Figure 2 represents the device diagramatically. The flow characteristics at the end of the nozzle have been

[1] Supported, in part, by a grant from the Office of Naval Research.

[2] A system composed of a number of units having variable distortion patterns in frequency and phase.

measured by means of an accelerometer and manometer, and recorded on a d.c. cathode-ray oscilloscope and high-speed camera. The value obtainable for the acceleration phase is 2 msec., and slightly more for the deceleration phase. When driving the arterial system, volumes ranging between 10 and 30 cc. are injected in total time periods ranging from 0.1–0.2 seconds. The position of this injected pulse in the cardiac cycle can be set by means of a variable delay circuit triggered from the R-wave of the electrocardiogram. Recording of the resulting pressure is obtained by means of a catheter lying in the ascending aorta adjacent to the injection nozzle. Orientation of the catheter tip and nozzle has been shown to be of no consequence since flow during these studies is turbulent. The catheter is coupled to a capacitance manometer, and the

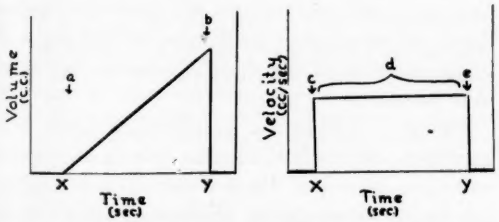


Fig. 1. *Curve A* represents graphically the relationship of volume entering the vessel to time. At *a* the solonoid valve opens. At *b* the solonoid valve closes. *Curve B* represents the first derivative of *curve A*. Thus, *curve A* represents velocity (cc/sec.) while *curve B* represents acceleration (cc/sec/sec.). Height of the rectangle represents the slope of *curve A; c* denotes acceleration; *d* denotes constant velocity; *e* denotes deceleration.

manometer output is recorded by means of either a d.c., two-beam, cathode-ray oscilloscope or a Brush ink recorder.[3] A second catheter, similarly coupled, lies in the femoral artery or abdominal aorta, thus providing data for pulse wave velocity measurements and the type of distortion occurring in the pulse wave in its travel down the vessel.

In the near future a detailed report concerning performance criteria for the driving device and recording equipment will be published. Suffice it to say that we are satisfied that we are able to record faithfully the pressure variations occurring within the vascular system.

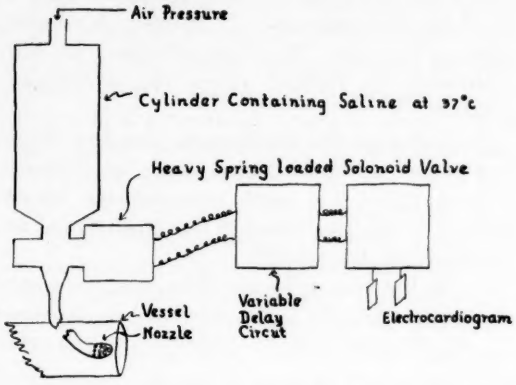


Fig. 2. SALINE is placed in the cylinder. Pressure varying from 25–150 lbs. has been applied to the fluid. The spring loaded valve opens very suddenly, and fluid flows from the nozzle. The valve is activated electrically. The R-wave of the electrocardiogram is amplified and sent through a variable delay circuit such that the volume pulse will occur at a pre-selected point in the cardiac cycle.

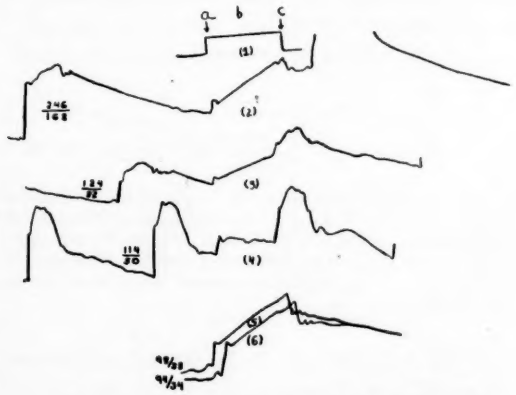


Fig. 3. *Curve 1* denotes the onset of flow into the aorta, *a*, the continued constant velocity of fluid entering the aorta *b*, and the cessation of flow into the aorta, *c*. *Curves 2 through 6* are the simultaneous pressure records taken from the aorta. *Curve 2* was seen following the injection of epinephrine; *curve 3*, the normal control; *curve 4*, following the intravascular administration of acetylcholine; *curves 5 and 6* are pressure pulse responses to the induced driving pulse and are not cardiac in origin. These latter two curves were seen following severe arterial hemorrhage.

RESULTS

Two types of dogs have been used in the work reported here. The first were those which were obviously puppies. The second were obviously old dogs. They varied in weight from 5–17 kg. Figure 3 demonstrates a typical series of curves, obtained in this case from a puppy. Several characteristics should be noted.

[3] The characteristics of these recording systems have been tested by means of pistonphone, transient analysis, and these methods were checked at the Underwater Sound Laboratory by the so-called free-field reciprocity method (2). The ink recorder coupled to the catheter manometer system was flat to between 30 and 40 c.p.s. With a cathode-ray oscilloscope, the range was extended to between 80 and 100 c.p.s. Using a steel needle, the range is extended to approximately 800 c.p.s. All three methods have been used. The Brush ink recorder and the cathode-ray oscilloscope have also been used simultaneously. Each has its advantage.

A. There is a sudden rise in pressure coincident with the onset of the injected pulse. This initial rise of pressure occurs during the period of acceleration (*A*, fig. 3) and at a time where there has been essentially no increase in volume within the aorta. This first pressure rise will henceforth be referred to as the *acceleration transient.*

B. Following the acceleration transient there is a period (*B*, fig. 3) during which the absolute volume within the aorta increases and hence there is a resulting pressure rise proportional to distensibility and run-off.

C. At the time when the flow into the aorta suddenly ceases there is a rapid fall in pressure (*C*, fig. 3) which is out of proportion to the run-off. This will henceforth be termed *deceleration transient.*

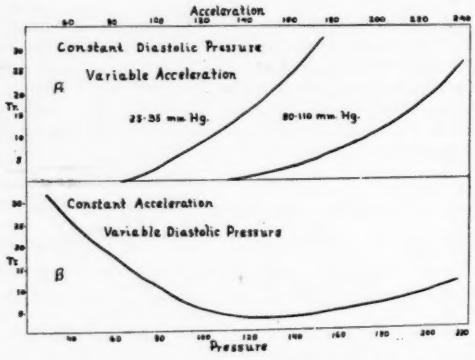


Fig. 4. *Curve A* represents the height of the acceleration transient, *Tr*, at two levels of diastolic pressure. The abscissa represents the acceleration of flow into the aorta. *Curve B* represents the magnitude of the acceleration transient, with constant acceleration, at various diastolic pressure levels. These curves are smoothed averages.

It became evident, from these studies, that the pressure pulse resulting from the volume pulse is due, not only to the absolute volume entering and leaving a segment of vessel and its distensibility, but also to the rate of change of velocity of fluid moving into the vessel and the state of the vessels themselves. Thus fluid enters the aorta, but the vessels can only begin to expand and fluid can only start to move within certain time limits. These limits of time and motion[4] may roughly be termed inertia characteristics. Fortunately we are able to study separately the effects of volume distensibility, i.e., dp/dv, and acceleration. If we increase the height of the square wave (*B*, fig. 3), (i.e., increase the magnitude of the initial acceleration of fluid into the aorta) we find that for a given diastolic pressure, the magnitude of the acceleration transient is likewise increased. At the same time the distensibility of this segment of aorta, as measured from the second part of the curve (*B*, fig. 3), has not changed. If now we maintain the value of the acceleration phase of the volume pulse constant and cause certain physiologic changes to occur within the animal, there are likewise certain changes in the magnitude of the acceleration transient. These effects occur when, for example, vascular tone or volume changes. These alterations which occur within the vessels themselves and which influence the magnitude of the acceleration transient have been lumped into a single term: the *stability state* of the vessels. The dynamics of this so-called stability state are complex, and attempts are now being made to define them. In our work so far it appears that the best single measure of the stability state is the diastolic pressure.

Figure 4 represents, graphically, these two variables which effect the magnitude of the acceleration transient. Dr. Lawton (3) has pointed out to me the similarity of these curves to those he has obtained from driven isolated strips of aorta.

The contour of the second part of the curve (*B*, fig. 3) on the other hand results from the familiar parameters: volume distensibility of the vessel and the ratio of volume entering and simultaneously leaving that segment of vessel. It is the distensibility (dp/dv) factor for which workers have been seeking a measure or, at best, some gauge of its effectiveness. Pulse wave velocity is the most commonly used index of distensibility.

From these studies it is evident that the arterial pressure pulse, as generated by cardiac action, results not only from the amount of blood the heart ejects (stroke volume), the distensibility of the system and run-off, but also upon the rate of change of velocity of the blood as the aortic valves open and upon the stability state of the vessels. This acceleration transient is a measure of the power developed by the heart and most certainly is similar to the force which produces the ballistocardiograph. I wish to reiterate that it is not directly related to the amount of blood the heart ejects, i.e., a large transient may occur with a very small stroke volume provided it enters the aorta rapidly enough. Its magnitude is further influenced by the stability state of the

[4] The resulting decrease in velocity energy appears as increased pressure energy.

vessels. The pressure rise, of course, would not be sustained if this acceleration of blood were not followed by an absolute volume change capable of distending the vessel. This is perhaps best demonstrated in figure 5, which shows a series of curves obtained from the injection of a similar pulse into the vena cava which is normally undistended, i.e., the walls are not round and under tension.

It is indeed interesting that under conditions where the heart tends to fail, as in experiments in which the thorax is open, this acceleration transient, which appears in the normal arterial pulse, tends to decrease in magnitude, and even disappear. This indicates that the heart is no longer able to muster enough force to accelerate

Fig. 5. *Curve A* represents type of pressure record obtained from the abdominal vena cava following injection of a pulse, in its normally undistended state. Pressure in the vein was approximately 10 mm. Hg. *Curve B* was obtained at approximately 15 mm. Hg as the vein wall was just distending. *Curve C* was obtained at a pressure of 25 mm. Hg; the vein in this case is distended.

the blood as previously. This occurs even before any other signs of failure occur, i.e., cardiac dilatation and elevated venous pressure. It is also apparent that the calculation of cardiac work should involve the use of this factor.

In summary then, the contour of the arterial pulse and thus the area enclosed by this contour and, therefore, the pressure energy available for the flow out of the arteries, is a resultant of several independent variables which must be accounted for in any procedure designed to measure stroke volume or hemodynamics utilizing pulse contour: *1*) The absolute volume injected into the vessels; *2*) the distensibility of the vessels; *3*) simultaneous run-off from the vessels; *4*) rate of change of velocity of blood entering and leaving the vessels; *5*) the 'stability state' of the vessels.

The pressure resulting from increased volume is superimposed upon whatever acceleration transient is generated; hence whenever the relative proportion of one to the other deviates from 'normal,' any simple contour method for calculating stroke volume not accounting for this will fail. If the arterial system is alinear in these properties, a further difficulty is encountered. This leads me to comment upon the use of pulse wave velocity, as an index of distensibility, in any procedure designed to calculate stroke volume from the arterial pulse contour. As vascular aging increases, certain changes may occur which alter markedly the effects of the loss of elasticity itself. These changes, which in a sense are compensatory, do not develop at the same rate or in the same areas in all persons. Two such characteristics of aging vessels are the increasing diameters of the aorta and the development of tortuosity throughout the system. The effect of these alterations is to increase the volume of the vessels transversely and longitudinally. These changes, of course, modify the pressure-volume relationships of decreased distensibility, and also

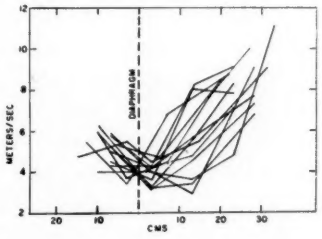


Fig. 6. GRAPH of pulse wave velocities in consecutive 10 cm. segments from ascending aorta to points noted in abdominal aorta. Note decrease in velocity about the axis of the diaphragm.

the significance of pulse wave velocity. Furthermore, for any individual arterial system, foreshortening of the vessels occurs when the mean arterial pressure falls. Preliminary observations of pulse wave velocity and pulse wave distortion in young and old dogs suggest that the systems behave differently quantitatively as well as qualitatively. Our observations in man show similar distortion differences with age. Another discrepancy will develop if tone changes occur in the arterial vessels. Again there is good, but preliminary, evidence that this occurs.

Even the measurement of pulse wave velocity is difficult and the possibility of error is considerable. Frequently the peripheral pulse has such a gradually developing wave front that it is difficult

to know exactly where to locate the onset of the wave. We have seen instances where we were sure that the pressure was falling and reflected waves were occurring which were equal to or greater than the initial inflection of the wave front, and the pulse wave was moved ahead, in time, of its actual velocity. Inasmuch as the pulse wave velocity is several meters per second and the lengths of vessel are relatively short, such errors make considerable difference in the final value.

If the above mentioned problems are not enough, the following two facts should discourage the use of pulse-wave velocity as a measure of distensibility. Figure 6 represents a series of values for pulse-wave velocity over the entire aorta of men for a number of individual ages (4). The alinearity of the system is obvious. We have also observed that, particularly in the young arteries, the velocity of the acceleration transient may be quite different from the distention wave.

Another question of importance, since an overwhelming proportion of work on this subject has been done on dogs, is whether the human arterial pulse has similar characteristics. We have measured arterial pulse curves from the root of the aorta to peripheral arteries in a large number of human subjects under various conditions. Dr. Schnabel has extended this series to include a greater variety of age and pathology. We have not, however, applied the method described above to man. A comparison of the contours and their changes under various conditions have demonstrated no differences between man and dog. Both man and dog show marked difference in the distortion of the central pulse as it is transmitted in the vessels of young and old.

I mentioned that the deriving method provides data which, I believe, can be used to calculate stroke volume. The method provides an absolute measure of the effects of acceleration on the system as well as the distensibility (dp/dv) and run-off occurring in the same segment of vessel into which the stroke volume of the heart is injected, and in which the recording catheter lies. These data can be obtained under any condition and at any time. Our work in this regard is preliminary and has not been extensively tested against other methods. Because of this and because the object of this panel was to discuss the Hamilton-Remington method and not to introduce new methods for the calculation of stroke-volume, I do not believe that additional space and time can justifiably be spent here in presenting these data in a proper manner. By way of an optimistic note, however, I would like to add that in a few cases we have compared stroke volumes calculated from the driving method and from the Hamilton-Remington method. The number is insufficient for statistical treatment but it appears that the two agree rather well under normal conditions. However, when the circulation has been altered so that the acceleration transient is disproportionately large, and especially when the pulse rate is high, they do not agree well. The Hamilton-Remington method gives figures for stroke volume which are considerably the larger of the two. The converse seems also to be true, i.e., with a failing heart and prolonged systolic ejection, their method provided values which were lower in volume.

REFERENCES

1. Peterson, Lysle H. *Federation Proc.* 11: 122, 1952.
2. Parnell, John. *The Evaluation of Manometer Systems.* Thesis, Univ. of Pennsylvania, Philadelphia
3. Lawton, Richard W. and Allen L. King. *Federation Proc.* 11: 90, 1952.
4. Schnabel, Truman G., Jr. *et al. Circulation* 5: 257, 1952.

VALIDITY OF THE PULSE CONTOUR METHOD FOR DETERMINING CARDIAC OUTPUT[1]

RUSSELL A. HUGGINS AND EDWIN L. SMITH

From the Department of Pharmacology, Baylor University College of Medicine, and the Department of Physiology, University of Texas School of Dentistry, Houston

MANY attempts have been made to circumvent the technically difficult and time-consuming direct Fick method for determining cardiac output. Another important limitation of the method is the requirement of a relatively steady state of the circulation for a period of at least 3 minutes and preferably 5 minutes. While this is not difficult to obtain under many circumstances, it presents, for example, a distinct problem in 'deteriorated' dogs where the removal of small quantities of blood may result in a significant decrease in cardiac output. More important, at least from the viewpoint of a pharmacologist interested in attempting an analysis of the mechanism and quantification of the cardiovascular action of drugs, is that many drugs have a relatively short duration of action. Direct Fick determinations of cardiac output under these circumstances are of questionable significance because the calculated blood flow is an averaged value for the period during which O_2 consumption is measured, and during this period there may have been significant changes.

The only other primary method of determining cardiac output in common use is by dye injection. This method has the great advantage over the direct Fick of reflecting the circulatory rate over a much shorter period of time, but like the direct Fick it has numerous technical difficulties. Therefore, some investigators have relied on such drastic procedures as cardiometry or have used a rotameter. However, pharmacologically speaking, mechanistic and quantitative inferences about changes in the cardiovascular system are frequently made from purely qualitative data.

It is not surprising then that numerous attempts have been made to develop a method of quantitating cardiac output that is technically simpler and more flexible. One of the directions that has been explored by many workers is the physical one in which the pressure values, pulse wave velocity and the pulse form are used to quantitate the cardiac stroke volume, but so far with questionable results. One of the reasons for the lack of success has been that the pulse wave velocity could never be a quantitative measure of total aortic volume uptake, a fact recently demonstrated by Remington, Hamilton, and Dow (1).

Later, Remington and Hamilton (2, 3) approached the problem from a different aspect by assuming it was feasible to ignore the absolute aortic capacity because it would be a quantity that varied with body surface area. They then collected their data into two tables, the first consisting of the volume uptake for each bed based only on $\Delta P/\Delta V$, and the second of transmission times to these beds. Using these pressure-volume and transmission-time tables, Bazett's method of calculating systolic drainage and a calculation of diastolic drainage, they constructed cardiac ejection curves from aortic pulse contours and demonstrated that these ejection contours satisfied the aortic pressure pulses. The next step was to utilize the aortic pulse contour for the calculation of stroke volume and compare the data with that obtained simultaneously from either the direct Fick or dye injection methods. In the original paper (4), an average difference of ± 8.2 per cent was reported, and in a subsequent paper (5) the average difference was reported as ± 10.1 per cent. The method was tested in our laboratory, using the direct Fick procedure as a basis of comparison on a small series of dogs (6) with an average difference of ± 12.2 per cent, whereas later in a small series of open chest dogs (7) the average difference was ± 12.5 per cent. Subsequently, Longino and Gregg (8) compared the stroke volume determined by a rotameter placed in the pulmonary artery and the pulse contour method, and reported an average difference of ± 12.0 per cent. In a recent abstract, Shadle *et al.* (9), using the dye dilution method

[1] This investigation was supported in part by a research grant from the National Heart Institute of the National Institutes of Health, Public Health Service.

as the basis of comparison, reported a range that was similar to that obtained by others but did not give their average difference. Finally, a paper by Duomarco *et al.* (10) reported results in which the left ventricular output calculated from the pulse contours was two or more times that of the measured right or left ventricular input, and in the appropriate paper will be made in presenting the results. Data of various kinds have been extracted and considered from approximately 250 dogs. The anesthetics used were either sodium barbital, 250 mg/kg. intravenously, or morphine sulfate, 10 mg/kg. subcutaneously followed in 15–20 minutes by sodium pentobarbital, 10 mg/

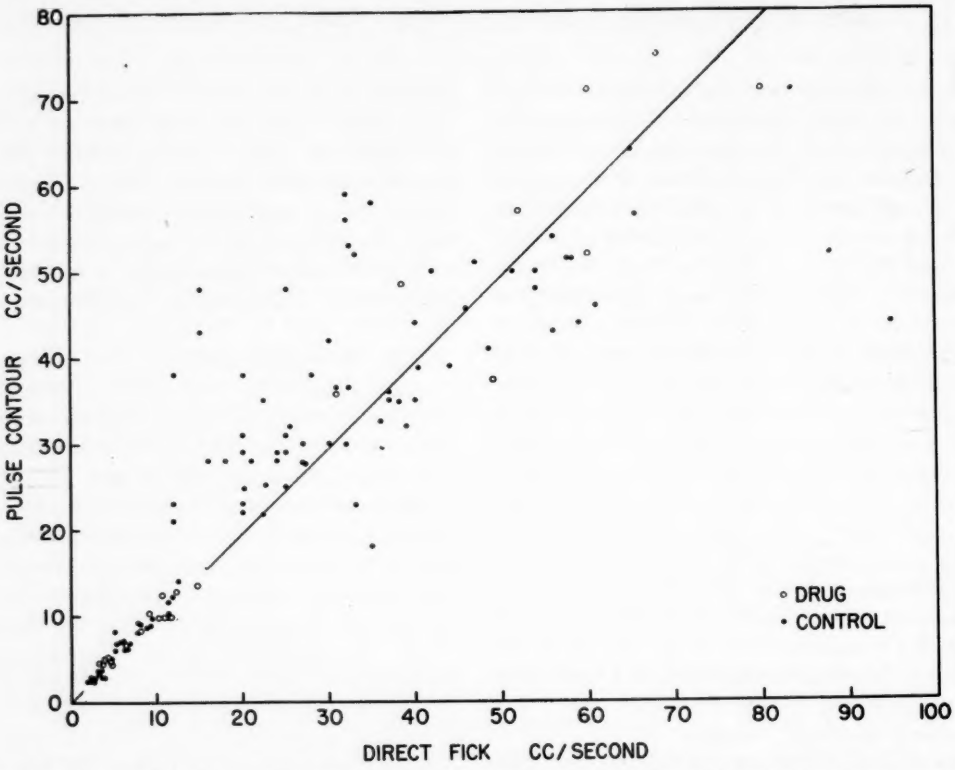


Fig. 1. Relationship between cardiac output determined simultaneously by direct Fick and pulse contour method in dogs subjected to minimal operative procedures. The solid line represents a perfect correlation.

some instances the directional changes were not indicated.

Considering the diversity of the methods used to test the validity of the pulse contour method, the errors of each, and the fact that they are based on different physiological principles, the reasonable agreement (except for the exception noted) found between the various methods suggests that the pulse contour method, although it may need modification, represents a distinct advance in the application of physical principles to the problem. This paper will present data on the validity of the pulse contour method under a variety of experimental circumstances.

METHODS

A number of procedures were used, most of which have been described before. Reference to

kg. intravenously. Various anticoagulants have been used, but during the last year and a half Paritol[2] has been used exclusively.

RESULTS

Figure 1 is a compilation of data comparing the aortic contour with the direct Fick method of determining cardiac output in relatively intact dogs. To minimize selection of the data all determinations have been plotted, irrespective of the fact that in some instances there was doubt concerning the validity of the Fick, or that in retrospect insufficient records of the pulse contour had been taken considering the variability in the mean pressure during the five-minute periods.

[2] Supplied through the courtesy of Dr. Joseph Seifter, Wyeth, Incorporated.

The only operative procedures were those necessary to obtain the blood samples and to record the pulse contour. A range of blood pressure was obtained by bleeding and by infusion of various drugs.

Under these conditions, the agreement between the two methods was good. When the data for the pulse contour method are examined, an average of 28.1 with a range of 3.0–76.0 cc/sec. is obtained; whereas for the direct Fick the average was 26.2 with a range of 2.0–95.0 cc/sec. Calculation of the standard difference between the two means gives the figure 0.74, indicating that there is no significant difference between the means. Examination of figure 1 suggests that above a cardiac output of 20 cc/sec. the pulse contour method tends to overestimate. One factor which could be concerned is blood pressure; but when a graph was constructed plotting the mean pressures between 40–116 mm. Hg against cardiac output, the results were inconclusive, particularly when the errors involved in the determinations were taken into consideration.

As a result of the reasonable agreement found between the direct Fick and contour methods, it was decided to utilize the latter method for more elaborate experiments. Briefly, the technical aspect of the experiment consisted in attaching a Dale-Schuster pump to the right auricular appendage so that a known and constant amount of blood was delivered to the right heart. Total venous return was measured with a large rotameter inserted into the superior and inferior venae cavae and emptying into a reservoir connected to the pump.

In the majority of instances, there was a large but variable discrepancy between the amount of blood being ejected from the left heart as calculated from the pulse contour and the known amount being pumped into the right heart (table 1, *exper. 1*). Further, in every instance where a significant difference existed, the pulse contour overestimated the true output. Occasionally, usually early in the experiment, the two methods agreed reasonably well; but eventually the calculated cardiac output from the pulse contour would become significantly greater than the known input (table 1, *exper. 2*). In several dogs, this was apparently precipitated by the administration of epinephrine (*exper. 2*) and once by mechanical manipulation.

Two types of response to epinephrine were observed under the conditions of this experiment. When agreement occurred between the pump input and the calculated output, administration of epinephrine resulted in a marked increase in the calculated output. In these instances, the calculated resistance increased but remained within reasonable limits (*exper. 2*). However, if the calculation of cardiac output was significantly above the input, injection of epinephrine sometimes resulted in the calculated outputs approaching the known input. This situation was characterized by a large increase in peripheral resistance (*exper. 1*). As the effects of epinephrine disappeared, the calculated output and pump input returned toward their original relationship. There developed, under the conditions of these experiments and in contrast to relatively intact dogs, a 'distortion factor' which caused the calculation of cardiac output to be significantly higher than the true output. It was impossible, *a posteri-*

TABLE 1. COMPARISON BETWEEN A KNOWN INPUT INTO RIGHT HEART WITH CALCULATION OF OUTPUT BY PULSE CONTOUR METHOD

PROCEDURE	MEAN PRESSURE	PUMP[1] FLOW	PUMP RESISTANCE	CONTOUR[1] FLOW	CONTOUR RESISTANCE
Exper. 1					
Control	130	19.4	5.7	36.6	3.0
Epinephrine .01 mg/kg.	215	19.4	10.0	23.3	19.4
Exper. 2					
Control	80	23.8	2.5	25.0	2.4
Epinephrine .01 mg/kg.	197	23.8	7.4	36.4	4.9

[1] cc/m²/sec.

ori, to decide which of the various instruments or operative procedures might be involved in the precipitation of this 'distortion factor'.

The first step in attempting to isolate this 'factor' was to determine whether or not opening the chest might be involved. Therefore, a series of 49 comparisons were made between the direct Fick and the pulse contour methods after opening the chest, and an agreement as good as that found in closed chest dogs was obtained (table 2). When a rotameter was inserted into the venous side of the circulation to measure total venous return, it was immediately apparent that a 'distortion factor' had been precipitated because the calculated output was again significantly higher than the measured return of blood into the right heart (table 3). That the rotameter is not necessary to the development of the 'factor' is shown

in table 2 where the only procedure was the constriction of the inferior vena cava to 3 or 5 mm. The calculated cardiac output was immediately elevated above the output determined by the Fick procedure. A detailed analysis of the changes in venous pressure which occurred under these circumstances is being published elsewhere (11).

Dibenamine (N,N-dibenzyl-β-chlorethylamine) was administered before and after the operative procedures without preventing the development of the 'distortion factor' or reversing it after its occurrence, because the condition of these animals was suggestive of shock (12) in spite of the relatively high mean pressures. No effect of Dibenamine was observed on the survival time of the animals in these experiments, which usually

TABLE 2. COMPARISON OF DIRECT FICK AND PULSE CONTOUR METHODS IN OPEN CHEST DOGS (CONTROLS) AND AFTER INFERIOR VENA CAVA IS CONSTRICTED TO 3 OR 5 MM.

	49 CONTROLS		76 IVC CONSTRICTIONS	
	Pm	C/F × 100	Pm	C/F × 100
Average	100.0	108.0	78.0	259.0
$\sqrt{\Sigma D^2/n}$	27.3	20.4	28.9	137.8

TABLE 3. COMPARISON OF RESULTS OBTAINED WITH ROTAMETER AND PULSE CONTOUR METHOD WITH RECALCULATION OF THE SAME DATA EXCLUDING VISCERAL BED

	Fc/Fr × 100	Pm	Fc − v/Fr × 100
Average	177.0	76.0	137.0
$\sqrt{\Sigma D^2/n}$	65.1	24.9	49.6

terminated in 2–4 hours. On the basis that a reflex might be involved, spinal anesthesia and transections were done without influencing the development of this 'factor' and under these circumstances the heart responded to Isuprel (1−(3′,4′dihydroxyphenyl)-2 isopropylaminoethanol) and stimulation of the vagus in the usual manner.

Returning to the data on the 'pump experiments,' we were impressed by the fact that throughout all the experiments there was never a time when the venous return exceeded the pump input into the right heart, and only very rarely, and then only for a brief time, were the two equal (table 4). In most of the experiments there was an almost constant amount of blood being lost from the circulation, necessitating a constant replacement of blood in the reservoir; in spite of this, at the termination of the experiment difficulty was often experienced in obtaining sufficient blood to calibrate the rotameter adequately. Therefore, an attempt was made to find the site of storage of this blood, in the hope that there might be some relationship between the distribution of this blood and the development of the 'distortion factor.'

The second step was to determine whether the blood deficit could be traced to the splanchnic area. In figure 2, data are presented for two dogs

TABLE 4. RELATIONSHIP BETWEEN CARDIAC INPUT AND VENOUS RETURN

PROCEDURE	PUMP[1] FLOW	VENOUS[1] RETURN
Control	23.8	19.0
Epinephrine .01 mg/kg	23.8	10.4
Control	19.4	16.9
Epinephrine .01 mg/kg	19.4	19.4

[1] cc./m²/sec.

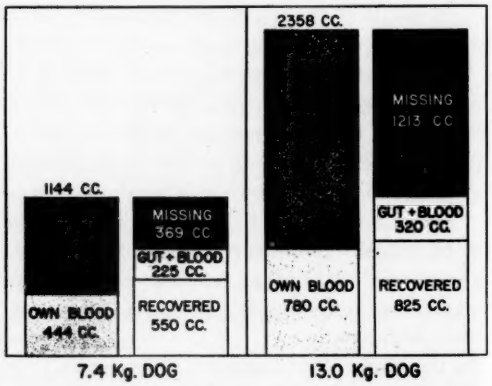


Fig. 2. RECOVERY of transfused blood in rotameter dogs. Data from two representative dogs of a series in which a rotameter has been inserted into the venae cavae and transfusions of blood from donor dogs made.

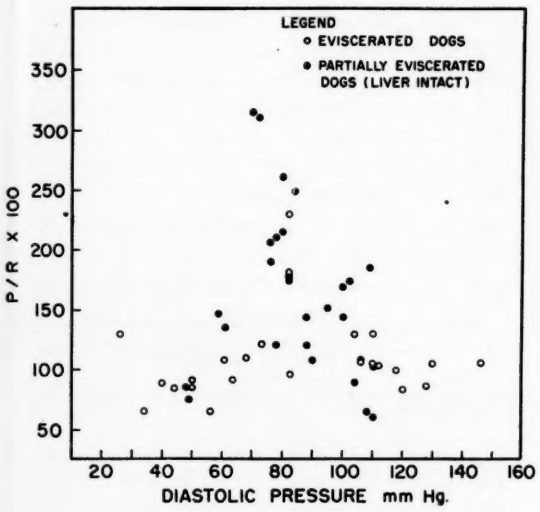
that are typical of this series. The initial blood volumes were estimated as 6 per cent of the body weight. For example, at the end of the experiment a 13 kg. dog after the addition of 1578 cc. of blood should have had an estimated blood volume of 2358 cc., but only 825 cc. were recovered. To simplify the estimation of the blood in the gut, the volume of the gut with its included blood was determined, obviously a gross overestimation of the blood volume. Even with this generous estimate, a total of 1213 cc. of blood is unaccounted for. Further, the data indicate that there has been no 'pooling' of blood in the gut,

and examination of the gut at the termination of these experiments showed that it was pale. The data then suggest that during the experiments in which the 'distortion factor' is operative there is a definite reduction in uptake and drainage of blood from the visceral area, with a consequent overestimation when cardiac output is calculated from the pulse contour. Also, the data clearly demonstrate that the 'pooling' or 'storage' of blood occurs in some area of the body other than the viscera. The effect this might have on the calculation of cardiac output from the pulse contour is not known; probably it will also cause an overestimation of output because of a reduction in drainage.

It should be possible to obtain data on the importance of these mechanisms by preparing eviscerate animals in which the rotameter is used to determine right ventricular input. In addition, a series was done in which the liver was kept intact. Because of the large number of comparisons between the two methods that can easily be made when the rotameter is used, no more than four points are graphed for any dogs, and the points for comparison were selected to represent a wide range in diastolic pressures produced by infusion of blood. Treating the data in this manner prevents too much weight being placed on a single experiment.

When the dogs were completely eviscerated,[3] good agreement, with the exception of one or two points, was found between the pulse contour calculation and the known right ventricular input (fig. 3), with an average of 102 per cent (contour × 100/rotameter) which compares favorably with an average of 108 per cent (table 2) found for open-chest and direct Fick dogs, and is in contrast to the uneviscerated rotameter series with an average of 177 per cent (table 3). The group in which the liver was kept intact is small (3 dogs), but the average is 162 per cent (fig. 3), a figure essentially the same as that of the uneviscerated rotameter series with an average of 177 per cent (table 3).

Examination of the data for the series of noneviscerated dogs in which the inferior vena cava was constricted and the rotameter inserted also showed that the pulse contour calculation reflected the directional changes, even when the



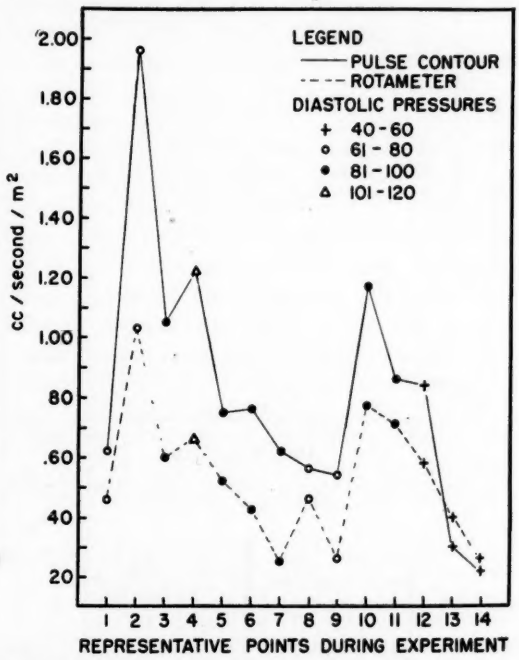
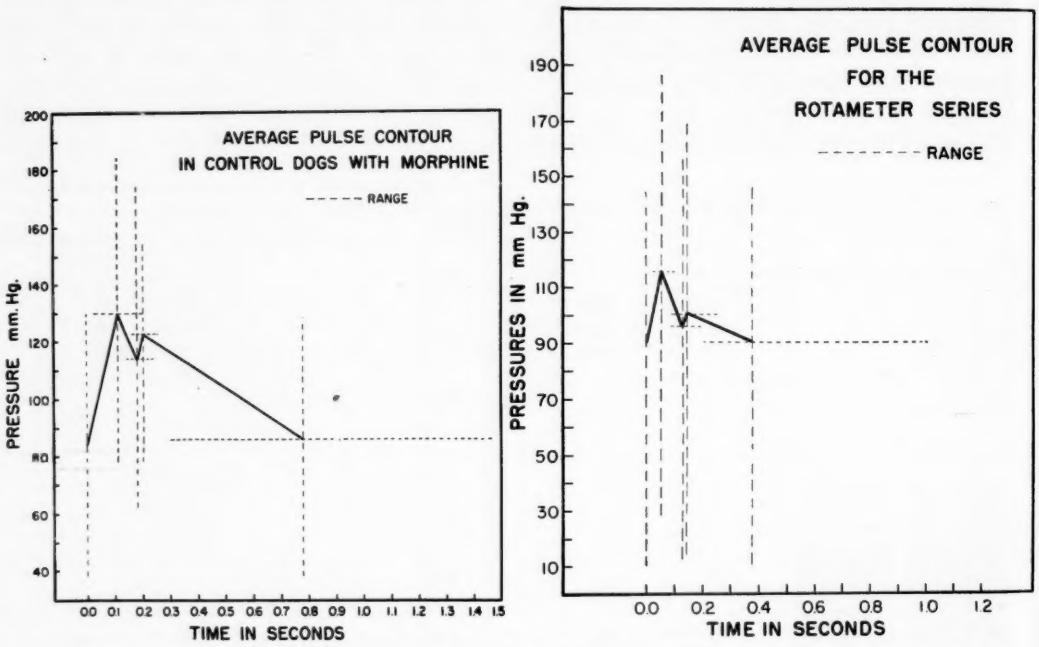


Fig. 3 (*left*). COMPARISON of simultaneous pulse contour and rotameter determinations of cardiac output plotted against diastolic pressure. 100 on the ordinate represents a perfect correlation.

Fig. 4 (*right*). DATA from partially eviscerated dog (liver intact) showing tendency for the pulse contour to follow directional changes in cardiac output in spite of 'distortion factor.'

[3] The eviscerations were performed for us by the kindness of Dr. Elliott B. Hay, Department of Surgery.

absolute flow was significantly lower. If the data which deviated most are plotted as successive points during the experiment for one of the dogs with an intact liver (fig. 4), the directional changes as calculated from the pulse contour reflect in all but one instance the changes measured with the rotameter, in spite of the criteria for detecting 'distortion contours,' our first procedure was to make a series of time and pressure measurements of pulse contours (fig. 5) for a group of 50 dogs anesthetized with morphine and pentobarbital. Aside from the anesthesia, the only other procedure on this group of dogs had been the insertion of a sound down a carotid for



Figs. 5 (*left*) and 6 (*right*). BROKEN HORIZONTAL LINES show the range in pressure and broken vertical lines the range in time.

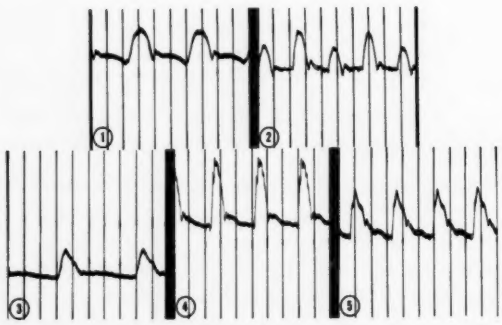


Fig. 7. VARIOUS TYPES of pulse contours seen in the 'rotameter series.' *Nos. 4* and *5* show the effect of Isuprel.

difference between the true blood flow and the calculated flow.

From a practical point of view, the important problem is whether or not it is possible by examination of a pulse contour to identify those that will not give a reasonably accurate determination of cardiac output. In attempting to establish the recording of pulse contours. A similar graph (fig. 6) was then made for all dogs in the 'rotameter series', a group characterized by 'distortion contours.' Important differences between the average contours for the two groups were evident, although there was some overlap in the ranges. The 'distortion contours' are characterized by a short cycle time, a steep anacrotic limb even at very low blood pressures, a marked drop in pressure after the anacrotic peak with the incisura occurring low on the dicrotic limb, occasionally even below diastolic pressure, and, finally, in many of the contours an almost flat diastolic slope (fig. 7).

Applying the above purely tentative empirical criteria to all the data, it was possible to separate the 'distorted' contours fairly successfully from the 'good' contours, but a residue of approximately 8 per cent fell in the uncertain class, although the pulse contour calculation of cardiac output was seriously in error. In these instances, though, it was possible by examination of the

pulse contours subsequently recorded in the experiment to realize that these represented potential 'distortion contours'. In our experiments, once the 'distortion contour' developed, there was never a reversion to a normal contour. However, the possibility that this 'distortion factor' may appear and disappear in relatively intact dogs, particularly when vasodilator drugs are used, has not been eliminated, but has never been observed in our fairly extensive series of experiments.

SUMMARY

In dogs subjected to a minimal amount of operative procedures, the pulse contour method for calculating cardiac output gave data in reasonable agreement with the direct Fick method.

Invariably, if the rotameter was inserted into the venae cavae or the inferior vena cava constricted to 5 or 3 mm., a 'distortion' of the pulse contour calculation of cardiac output occurred when compared to the rotameter or direct Fick methods. The 'distortion' consisted of the calculated output from the pulse contour being definitely, by a variable amount, above the cardiac output determined by other methods. In spite of this 'distortion factor,' the pulse contour calculations continued to show directional changes in cardiac output.

Various attempts were made to discover the mechanism of the 'distortion factor.' Evidence was obtained indicating that simultaneous with the development of the 'distortion factor' there was a marked reduction of uptake and drainage from the visceral area. When dogs were eviscerated and comparisons made between simultaneous rotameter and pulse contour calculations of cardiac output, the two methods were in good agreement; but if all the viscera were removed except the liver, the agreement was more variable, with a 'distortion factor' present in some instances. Recalculation of the 'rotameter series', omitting the uptake and drainage for the visceral bed, greatly improved the agreement between the rotameter and pulse contour calculations, but not sufficiently to indicate that uptake and drainage from the visceral area was solely responsible for the 'distortion factor.'

By plotting a large series of the 'distorted contours', suggestive criteria for differentiating them from normal contours were obtained. These were: short cycle time, steep anacrotic limb even at low blood pressures, a marked drop in pressure after the anacrotic peak with the incisura occurring low on the dicrotic limb or occasionally even below diastolic pressure, and, finally, in many instances an almost flat diastolic slope. These contours may occur within the normal range of blood pressure. Application of these empirical criteria to all our data separated all but 8 per cent of the 'distorted contours'. Further examination of this 8 per cent indicated that analysis of subsequent contours revealed that 'distortion contours' eventually developed.

REFERENCES

1. Remington, J. W., W. F. Hamilton and P. Dow. *Am. J. Physiol.* 144: 536, 1945.
2. Remington, J. W. and W. F. Hamilton. *Am. J. Physiol.* 144: 546, 1945.
3. Remington, J. W. and W. F. Hamilton. *Am. J. Physiol.* 148: 25, 1947.
4. Hamilton, W. F. and J. W. Remington. *Am. J. Physiol.* 148: 14, 1947.
5. Remington, J. W., W. F. Hamilton, N. C. Wheeler and W. F. Hamilton, Jr. *Am. J. Physiol.* 159: 379, 1949.
6. Huggins, R. A., C. A. Handley and M. La Forge. *Proc. Soc. Exper. Biol. & Med.* 68: 543, 1948.
7. Huggins, R. A., E. L. Smith and M. A. Sinclair. *Am. J. Physiol.* 160: 183, 1950.
8. Longino, F. and D. Gregg. *Am. J. Physiol.* 167: 721, 1951.
9. Shadle, O. W., T. B. Ferguson and D. F. Gregg. *Federation Proc.* 11: 145, 1952.
10. Duomarco, J. L., W. H. Dillon and C. J. Wiggers. *Am. J. Physiol.* 154: 290, 1948.
11. Smith, E. L., R. A. Huggins, R. W. Randall and G. A. Jeffery. *Texas Reports in Biol. & Med.* In press.
12. Wiggers, H. C., R. C. Ingraham, F. Roemhild and H. Goldberg. *Am. J. Physiol.* 153: 511, 1948.

AMERICAN INSTITUTE OF NUTRITION

Symposium on Geriatric Nutrition[1]

Chairman: C. M. McCay

INTAKES AND RETENTIONS OF NITROGEN, CALCIUM AND PHOSPHORUS BY 136 WOMEN BETWEEN 30 AND 85 YEARS OF AGE[2]

MARGARET A. OHLSON,[3] WILMA D. BREWER, LOIS JACKSON, PEARL P. SWANSON,[4] P. HARRIETT ROBERTS, MARGARET MANGEL,[5] RUTH M. LEVERTON,[6] MARILYN CHALOUPKA, MARY ROSE GRAM, MAY S. REYNOLDS[7] AND RUTH LUTZ

ANIMAL studies suggest that food may profoundly affect the quality of adult vitality and the time of development of characteristics of aging (1, 2). Today, a larger number of Americans are living into the older age brackets than formerly. The social problems of an aging population and the value to society of prolonging the potential work capacity of the individual are accepted, but few investigations of aging human populations have been made.

The present study[8] was initiated in 1946 for the purpose of obtaining data on food intake and nutritional status of women who had passed the 30th birthday. It was elected to study the aging rather than the aged since it seems probable that the role of nutrition in the health of the adult may be in the delay of those cellular changes which are found in old age. Data for the utilization of nitrogen, calcium and phosphorus of self-selected diets by 136 women who acted as subjects for balance studies for a period of at least a week will be reported in this paper. The experimental approach has been that of the ecologist who attempts to observe an individual within his environment and with a minimum of disturbance of his behavior. Data which are presented should be considered as a description of the performance of a population segment rather than an attempt to define minimum requirements for the three nutrients in question.

DESCRIPTION OF THE SAMPLE

The distribution of the 136 subjects by states and age is given in table 1. The sample is essentially a selected one. Socio-economic status was evaluated by the investigator in each area. Details for socio-economic ratings will be published later.

The Iowa investigators worked from an area sample which was prepared by the Statistical Laboratory of Iowa State College. The area sampling units used comprised clusters of households. The sampling units were selected in the city of Ames and the two rural townships in which Ames is located, Franklin and Washington townships, Story County, Iowa. Personal interviews with occupants of the homes found in the clusters revealed whether any woman over 30 years of age living in each home was willing to take part in the project. No woman was excluded from the study on the basis of health, although many excluded themselves at the interview because of poor health. Thirty-six of the households could be considered of average socio-economic status or above, 8 were below average. Forty-three women had been married and 39 had borne one or more children.

The Nebraska sample was recruited largely from among the University community. Fourteen

[1] New York City, April 16, 1952.

[2] Contribution No. 4, Subproject 1, "The Nutritional Status and Dietary Needs of Older People" of the North Central Region Cooperative Project NC-5 "Nutritional Status and Dietary Needs of Population Groups."

[3] Authorized for publication as Journal Article No. 1363 of the Michigan Agricultural Experiment Station.

[4] Journal paper No. J-2102 of the Iowa Agricultural Experiment Station, Project No. 1028.

[5] Journal paper No. 1312, approved by the Director of the Missouri Agricultural Experiment Station.

[6] Published with the approval of the Director as paper No. 509, Journal Series, Nebraska Agricultural Experiment Station.

[7] Published with the approval of the directors of the Wisconsin Agricultural Experiment Station.

[8] In cooperation with the Bureau of Human Nutrition and Home Economics, U. S. Department of Agriculture.

families were of average socio-economic status and 38 families of upper socio-economic status. Forty-three of the women studied had been married and 31 had borne children.

Samples from Michigan and Missouri were of similar composition in that individuals were recruited from the communities in which the respective experiment stations were located, without emphasis on socio-economic status and with no attempt at randomization for locating individuals. About a third of the subjects from Michigan and Missouri were from low socio-economic groups, though none of them were indigent. The remainder of the sample was at least comfortably situated with respect to buying power. All but six of the women in these states had been married and had borne one or more children.

The Wisconsin sample comprised wives of graduate students and instructors on the University campus. Socio-economic ratings were average or above. All women had borne children.

Women from all States were active, participating members of the household and many of them had a part in community activities. All were physically able and otherwise competent to carry out the procedures of weighing the diet, making the necessary collections of excretions and of assisting in the keeping of detailed records.

Preliminary physical examination and medical history excluded the presence of active, uncontrolled disease known to affect the utilization of food or seriously limit physical activity. Studies of the formed elements of the blood, certain measures of blood chemistry, and basal metabolism tests made by the various laboratories were within generally accepted standards and will not be discussed here.

Since there are degrees of vitality represented within an active and essentially well group of persons, a health rating was devised which included pertinent aspects of both the medical history and the physical examination (table 2).

TABLE 1. DISTRIBUTION OF SUBJECTS ACCORDING TO STATES

STATE	AGE RANGE—YEARS						
	30–39	40–49	50–59	60–69	70–79	80–89	Total
Nebraska	14	15	13	8	2		52
Iowa	6	14	10	8	5	1	44
Wisconsin	5	2					7
Missouri		2	10	3			15
Michigan		1	7	6	4		18
Total	25	34	40	25	11	1	136

TABLE 2. DISTRIBUTION OF SUBJECTS ACCORDING TO HEALTH RATING

AGE RANGE	NO. SUBJECTS	HEALTH RATING		
		Good	Fair	Poor
yrs.				
30–39	25	16	7	2
40–49	34	9	18	7
50–59	40	21	18	1
60–69	25	12	12	1
70–79	11	4	6	1
80–89	1	1		
Total	136	63	61	12

TABLE 3. PERCENTAGE OF OVERWEIGHT AND UNDERWEIGHT INDIVIDUALS AMONG SUBJECTS[1]

AGE RANGE	NO. SUBJECTS	UNDERWEIGHT	OVERWEIGHT	
		Over 15%	15–29%	Over 30%
yrs.				
30–39	25	4		4
40–49	34	9	18	15
50–59	40		37	8
60–69	25		40	28
70–79	11		18	10
80–89	1			
Entire series	136	3	24	13

[1] Sherman, H. C. (3).

This is admittedly a subjective evaluation and may give undue weight to past illnesses from which good recovery has been made. It is of some interest that 63, or nearly half of the subjects, were rated as being in good health; 36 of the 61 women classified as being in fair health were in the decades during which symptoms of the menopause are prominent, as were 8 of the 12 individuals considered to be in poor health. Of these 44 cases, 22 were judged to be in fair or poor health because of symptoms and medical findings related to the circulatory or nervous system which, at least in part, could be related to menopausal changes.

The most frequently occurring nutritional defect was the deviation of weight from medical standards. Table 3 records the percentage of cases in each age group which were more than 15 per cent underweight, 15–29 per cent and over 30 per cent overweight, according to the stand-

ards of Wood (3) which assume weight at age 30 to be suitable for the following decades. Weight judgments in this case were made without attempt to assess body build. Two ranges of overweight are given, since medico-actuarial weight tables assume some weight increase with age. If the latter criterion had been used, most of the cases in the 15–29 per cent overweight column could be considered to be of normal weight. The persistence of excessive overweight in 28 per cent of the women in the 60's and 10 per cent of the women over 70 years is of interest as the group probably represents a sample of vigorous women of this age range, since the physically infirm excluded themselves.

It is obvious that the total sample is weighted, since nearly half of the women were considered to be in good health and 76 per cent were considered to have the buying power and social status of an average family or better. It is probable that all cases could have purchased an adequate diet but that this would require better than average management practices in 24 per cent of the cases.

EXPERIMENTAL PROCEDURES

In all cases the woman was studied in her own home and was paid for her participation. Laboratory staff members instructed each woman in the techniques of weighing food and collecting and saving excreta, and visited the home at least once a day during the actual period of study. Staff members were available at other times as needed to insure adequate experimental supervision. Since one objective of the experiment was to influence the women's usual habits as little as possible, both experience and finesse on the part of the research personnel were required so that supervision was adequate without obstructing the women's usual routines. Details of collection periods are found in table 4. Although no preliminary period was used in certain States, subjects were instructed before the beginning of the balance period.

Diets were completely self-selected both as to the kind and amount of food eaten and the time of day at which meals were taken. The woman's portion of each food was weighed on a spring balance at the time that the family meal was served. Uneaten food was weighed at the end of the meal and subtracted from the total and an aliquot of each food eaten was set aside for the laboratory. Measured amounts of tap water were used for drinking and for such beverages as coffee and tea and suitable aliquots added to the food samples. Food composites were blended and frozen or were converted into brown digests by the method of Stearns (4) and stored until analyzed.

Carmine was used to mark the feces. The fecal collections for several days were blended and sampled for storage by freezing or as a brown digest. Urine was analyzed at once or acidified and stored. Nitrogen was determined by the macro-Kjeldahl method, calcium by the McCrudden method described by Stearns (4), and phosphorus according to the method of Fiske and Subbarow (5).

RESULTS

There were differences between States in sample and in experimental plans used. For this reason the data were examined first by State groups and the correlation coefficients of intake and retention were computed for nitrogen, calcium and phosphorus for each locality and age group. The coefficients were similar within each decade so that the locality differences in the behavior of subjects to a given intake of any of the three nutrients could be attributed to sampling variation. However, there were differences in the responses of each age group in the various localities. Thus, data to be presented will be combined for localities but the age groups will be considered separately. The data are presented as mean daily intakes and retentions. The influence of body size will be considered at a later time.

TABLE 4. SAMPLING PLANS

STATE	PERIODS OF WEIGHED FOOD INTAKE		
	Preliminary	Balance	Post-Balance
	days	*days*	*days*
Iowa	0	7[1]	0
Michigan	7–10	10	45–60
Missouri	0	10	21
Nebraska	0	7	0
Wisconsin	10	10	10

[1] 5 days only reported for 3 subjects.

Nitrogen: The mean daily intake and retention of nitrogen for each decade are presented in table 5. The measure of variation is the standard error of the mean. Mean intakes of calories were computed (6–8) and are included in this table. Fifty-eight per cent of the group were in negative nitrogen equilibrium.

Nitrogen intakes ranged from 3.39–16.86 gm. per day and both extremes of the range were found in the 50–59 age group. Mean nitrogen intake decreased with each decade. To test the

significance of the decreases, a regression of nitrogen on age was computed and the slope of the regression line tested against the hypothesis that there was no change in intake of nitrogen with age. The value for t was 3.01 with less than one chance in a hundred of a chance result. Mean calorie intakes also decreased significantly with age (t = 5.55).

The coefficients of correlation of intake and retention of nitrogen for the respective age groups were 0.60, 0.53, 0.61, 0.43 and 0.35 (table 5).

Figure 1 shows the relation of mean calorie and nitrogen intakes to nitrogen retentions. An examination of the distribution of retentions by age for increasing mean intakes of nitrogen indicated that the age groups 40–49 and 50–59 could be combined, as could all ages over 60. This was done to increase the clarity of the figure.

When 1500 calories or less were eaten, daily mean negative nitrogen retentions occurred even though an average of 8–10 gm. nitrogen were ingested. The age groups below 40 and over 60

TABLE 5. MEAN INTAKES AND RETENTIONS OF NITROGEN OF WOMEN, SUCCESSIVE DECADES

AGE	NO.	MEAN INTAKE	MEAN RETENTION	NO. NEGATIVE RETENTIONS	COEFFICIENTS OF CORRELATION[2]	CAL/DAY
		mg/day	*gm/day*			
30–39	25	10.85 ± 0.37[1]	0.08 ± 0.24[1]	13	0.60**	1915
40–49	34	10.64 ± 0.30	−0.18 ± 0.30	19	0.53**	1795
50–59	40	9.93 ± 0.37	−0.47 ± 0.23	24	0.61**	1752
60–69	25	9.12 ± 0.37	−0.46 ± 0.24	17	0.43*	1630
70–79	11	8.15 ± 0.51	−0.32 ± 0.36	6	**0.35**	1575
80–89	1	9.08	1.33			1531

[1] Standard error of the mean. [2] * = $P<0.05$; ** = $P<0.01$.

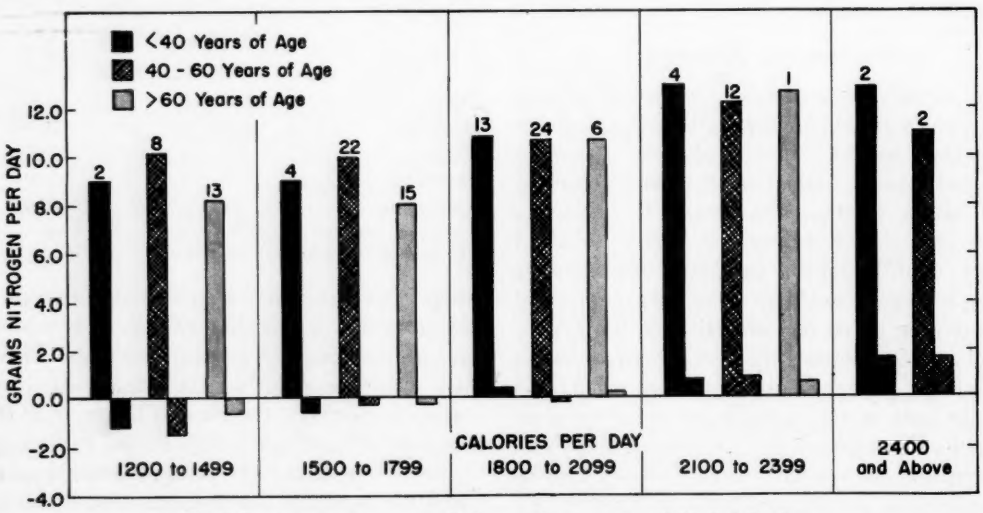


Fig. 1. INTAKE AND RETENTION of nitrogen of women at graded calorie intakes.

The probability that intake and retention of nitrogen are correlated is high ($P < 0.01$) for the ages between 30 and 59; between 60 and 69 there is only one chance in twenty of a high correlation, and after the age of 70 there was no correlation between intake and retention of nitrogen.

As might be expected, the intake of calories and protein also was highly correlated except in the 70–79 year old group. The correlation coefficient for the entire series was 0.68. It was 0.70 for ages 30–39, and 0.64, 0.60, 0.73 and 0.46 for the successive age groups.

years show a regular increase of mean nitrogen intake with each increase in calories which is reflected in an equally regular increase in the mean retention. This response might be expected for a group of adults who show no overt signs of nitrogen depletion and who probaby exist in a range of nitrogen nutrition somewhat below saturation of nitrogen stores. Approximately 1800 calories per day were required to insure nitrogen equilibrium.

The interpretation of the data for the years 40–59 is less clear. Again nitrogen losses occurred when less than 1500 calories per day were eaten

and retentions occurred when calories were increased above 2100 per day. Between 1500 and 2100 calories, mean retentions suggested nitrogen equilibrium although there were individual cases with large gains or losses of nitrogen at each calorie intake. The interrelationship of calories and nitrogen in promoting nitrogen balance are obvious on these self-selected diets, but certain individuals within the 40–59 age range were able to maintain body stores on limited intakes of both nitrogen and calories while others seemed to be unduly extravagant in their requirements.

Intakes of protein required for equilibrium under the conditions of calorie intake and self-selected diet of this experiment were computed by solving the equation $Y = a + bX$ for intake at zero retentions of nitrogen and multiplying by the factor 6.25. The estimates were: 66 gm. at ages 30–39, 69 gm. at 40–49, 70 gm. at 50–59, 67 gm. at 60–69 and 59 gm. after 70. The higher predicted requirement for ages 40–49 and 50–59 lend further support to the concept of stress during these years. It is of interest that the predicted protein required for equilibrium remains high until age 70.

Calcium: Table 6 gives the mean intakes and retentions of calcium for each age group with the respective standard errors. Individual intakes ranged from 0.31–1.42 gm. per day excluding one case who ingested a variety of dietary supplements providing more than 2 gm. calcium per day. Although 54 per cent of the cases were in negative calcium balance, only 38, or 28 per cent, excreted 0.1 gm. or more than ingested.

The calcium intake did not change significantly until after the age of 70. The decrease in intake from the 60–69 year series to the 70–79 year group is a significant one in spite of limited numbers of subjects in the latter group. On the basis of mean retentions, each age group may be considered to be in calcium equilibrium, though individuals within the age groups gained or lost appreciable amounts of calcium.

The distribution of the retentions of each decade for graded intakes of calcium is shown in figure 2. The correlation coefficient of intake and retention for the entire series was 0.52 and is highly significant. However, correlation coefficients for each decade were of doubtful significance for the 30–39 year old group ($P < .05$) and were not significant after age 70; coefficients were highly significant for the three intermediary decades ($P < .01$). The data suggest that there is a wide range of calcium intake within which individuals under 40 and over 70 years of age tend to establish calcium equilibrium, and that calcium was stored when more than a gram per day was ingested.

Negative calcium balances in the 60–69 year group decreased as the intake increased from 0.37–0.7 gm. per day. The average equilibrium was attained between 0.7 and 0.9 gm. per day. The metabolism of the 40–59 year old groups is less easily interpreted. Subjects between 50 and 59 years of age lost significant mean amounts of calcium until the intake exceeded 0.9 gm. per day. Five-eighths of this group ingested less than 0.9 gm. of the mineral per day. When intakes exceeded 0.9 gm. per day, this age group retained the most calcium of any age classification. The metabolism of the 40–49 age group is intermediate when compared with the immediately preceding or following decades.

Further evidence of possible stress associated with the ages from 40 through 65 is found in the data for urinary calcium which are shown in table 7. Although intakes in the decade groups from 40 through 69 were not greatly different, and were smaller than intakes in the 30–39 age group, both the mean and the percentage of intake excreted in the urine increased to a peak in the 50–59 decade and then decreased. The high percentage excretion in the 70–79 age group was

TABLE 6. MEAN INTAKE AND RETENTIONS OF CALCIUM OF WOMEN, SUCCESSIVE DECADES

AGE	NO.	MEAN INTAKE	MEAN RETENTION	NO. NEGATIVE RETENTIONS	COEFFICIENTS OF CORRELATION[2]
		gm/day	*gm/day*		
30–39	25	0.95 ± 0.06[1]	0.01 ± 0.03[1]	10	0.43*
40–49	34	0.84 ± 0.05	−0.01 ± 0.03	20	0.44**
50–59	39	0.83 ± 0.05	0.00 ± 0.03	21	0.61**
60–69	25	0.82 ± 0.07	−0.03 ± 0.03	14	0.63**
70–79	11	0.65 ± 0.05	−0.02 ± 0.03	8	0.28
80–89	1	0.98	−0.05		

[1] Standard error of the mean. [2] * = $P<0.05$; ** = $P<0.01$.

due to an excessively high urinary calcium of one subject. There were a total of 16 subjects in the entire series who excreted more than 0.240 gm. calcium in the urine per day, which, from the study of Knapp (9) might be considered to be a high urinary loss for the intakes which were observed. Two of these subjects, one in her 30's and the other 63 years of age, ingested 1.34 and 2.09 gm. per day and excreted 0.240 and 0.292 gm. respectively. The remaining 14 subjects were distributed as follows: 2 in the 40–49 group, 9 in the 50–59 group, 2 in the 60–69 group and one over 70. The 13 subjects who ranged in age from 46–65 years averaged 0.284 gm. urinary calcium per day, which represented 36 per cent of the intake (25–59 per cent). Intakes ranged from 0.43–1.13 gm. with an average of 0.83 gm. per day. Six of these 13 subjects were more than 30 per cent overweight and 4 were receiving some form of hormone therapy. Nine of the 13 were rated to be in fair or poor general health.

The predicted calcium intakes at which equilibrium would be expected under the conditions of this experiment were computed (using the equation of a straight line) and were: 0.88, 0.87, 0.83, 0.92 and 0.73 gm. per day for each age group from 30–39 to 70–79 respectively. When diets are self-selected, there would appear to be no decrease in the requirement of calcium of the average vigorous woman until after the age of 70. This is in keeping with data reported previously from two of these laboratories for women 50–80 years of age (10, 11).

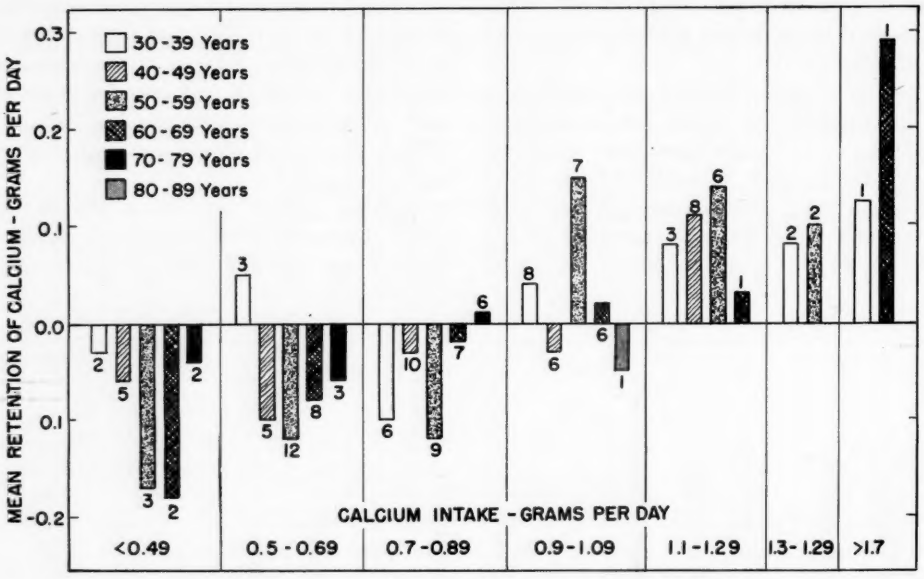


Fig. 2. MEAN RETENTION of calcium of women at graded calcium intakes.

TABLE 7. MEAN CALCIUM INTAKE AND URINARY EXCRETION OF CALCIUM

AGE	NO.	MEAN INTAKE	MEAN CA. IN URINE	
		gm/day	*gm/day*	*% intake*
30–39	25	0.95	0.149	15.7
40–49	34	0.84	0.160	19.0
50–59	39	0.83	0.185	22.3
60–69	25	0.82	0.151	18.4
70–79	11	0.65	0.143	22.0
80–89	1	0.98	0.133	13.6

It is interesting and somewhat unexpected that the intake of calcium also was correlated with the intake of calories ($r = 0.49, P = <.01$).

Phosphorus: Table 8 shows intakes and retentions of phosphorus for each group. Phosphorus intakes ranged from 0.49–1.70 gm. per day. The gradual decrease in phosphorus intakes with increasing age follows the pattern of response seen in the nitrogen studies and is highly significant ($t = 3.35$). The intake of phosphorus also is highly correlated with the intake of calories ($r = 0.66$ for the entire series) as was true of the intake of nitrogen and calcium.

Intake and retention of phosphorus were significantly correlated for the age groups from 40 through 69 ($r = 0.49, 0.54, 0.55$ respectively, $P < 0.01$ in all cases). There was no correlation between intake and retention in the 30–39 year group and only one chance in twenty of a significant correlation after age 70 ($r = .63$). This response is similar to that observed for calcium.

Figure 3 shows the grouping of retentions at each decade for graded intakes of phosphorus. Individual subjects less than 40 and over 70 years would appear to be able to maintain phosphorus equilibrium at a wide range of intakes up to 1.3 gm. At higher intakes, the younger group retained phosphorus in amounts increasing with each increment of intake. None of the cases of retention. The prediction for phosphorus intakes required for zero equilibrium (using the equation of a straight line) yielded values of 1.25, 1.32, 1.42, 1.51 and 1.13 gm. per day for ages 30–39 to 70–79 respectively. The apparently high requirement from 50–69 is of interest as is the apparent decrease in phosphorus requirement after age 70.

TABLE 8. MEAN INTAKES AND RETENTIONS OF PHOSPHORUS OF WOMEN, SUCCESSIVE DECADES

AGE	NO.	MEAN INTAKE	MEAN RETENTION	NO. NEGATIVE RETENTIONS	COEFFICIENTS OF CORRELATION[2]
		gm/day	*gm/day*		
30–39	25	1.24 ± 0.05[1]	0.00 ± 0.03[1]	11	0.30
40–49	34	1.12 ± 0.04	−0.07 ± 0.03	23	0.49**
50–59	40	1.08 ± 0.04	−0.13 ± 0.03	31	0.54**
60–69	25	1.05 ± 0.06	−0.11 ± 0.03	19	0.55**
70–79	11	0.91 ± 0.05	−0.06 ± 0.02	7	0.63**
80–89	1	1.11	0.10		

[1] Standard error of the mean. [2] * = $P<0.05$; ** = $P<0.01$..

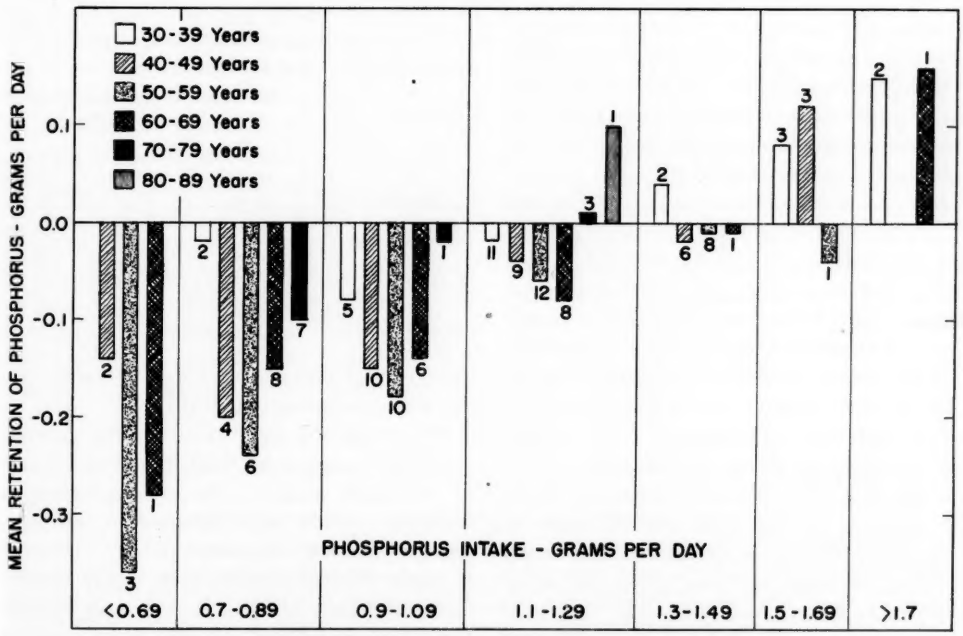


Fig. 3. MEAN RETENTION of phosphorus of women at graded phosphorus intakes.

older group ingested more than 1.3 gm. of phosphorus per day.

Phosphorus losses tended to be large for the ages 40–69 and to decrease with each increment in intake until 1 gm. or more per day was ingested. The greatest losses are found in the age period 50–59. A consistent trend in retentions for this decade is not apparent until intakes exceed 1.49 gm. per day. Since few subjects ingested this much phosphorus on self-selected diets, no significance can be attached to the few

DISCUSSION

There are few published studies which can be compared directly with this work. Previous papers from these laboratories (11, 12) have reported similar data on 28 women 50–77 years of age. The results are in essential agreement with the data presented in this paper when allowance is made for differences in age distribution of the samples.

Kountz (13) has reported nitrogen balances

on "a number of older people" and found that 1 gm. protein per kilogram body weight was required for equilibrium and that infection and other illnesses tended to produce negative retentions. If protein is calculated on the basis of absolute weights, the mean requirements ranging from 57–70 gm. per day for the decade groups of this experiment would represent approximately one gram per kilogram in the younger women and somewhat less than a gram per kilogram for women over 70. If computed on the basis of a 'normal' adult weight, the estimate represents 1–1.3 gm. per kilogram.

While these appear to be high requirements for nitrogen, there are several factors which would tend to increase nitrogen wastage. Caloric intakes averaged less than 2000 per day for all age groups. Food was entirely self-selected and the distribution of kind and amount of protein among the meals and the time of day at which meals were eaten were both variable.

Leverton and Gram (14) have demonstrated that shifting one portion of good quality protein from one meal to another can change the apparent requirement of nitrogen for equilibrium. Low calorie intakes also would be expected to increase protein requirements (15).

Bricker and her coworkers (16) and Hegsted *et al.* (17) have reported nitrogen equilibrium at intakes less than half of those recorded in this paper. In the former study, caloric intakes were a fourth to a third higher than those found in this series. In both studies, the nitrogen of the diet was fed on a strictly prescribed mealtime schedule and the women subjects were all young adults whose performance is comparable only to that of the youngest age group of this study.

The role of food in the social structure of the family group is an important part of many of the irregular eating practices observed. The heavy noon dinner on Sunday in contrast to the usual weekly dinner hour and the tendency to serve large portions of protein food at a special meal and a heavily 'extended' protein dish on other days are characteristics of an average household in which the woman shares. As long as food is available and they have the buying power, the average Middle Western family with its traditions of personal liberties is not apt to give up these customs, which have a social basis.

The practical implications of these observations are important in that women approaching middle age are quite uniformly advised to reduce calories if there is any tendency to weight gains. Such advice is all too often given without exploration of the eating practices of the woman and it is assumed that she will have the knowledge and judgment to select food high in specific nutrients. In a characteristic Middle Western diet the dilution of specific nutrients with fat and refined carbohydrates is such that careful selection is required as calories are reduced, or nutrients will be reduced proportionately. Few untrained women can make these selections without specific instruction. Inadequate breakfasts and poorly chosen lunches undoubtedly result in nitrogen wastage even when the dinner meal is large and total intake of protein would seem to be adequate. It is suggested, though experimental proof is inadequate, that the daily variables characteristic of freely chosen meal patterns both as to kinds and amounts of food served and eating schedules, may preclude the type of adjustment to low intakes of protein and perhaps

TABLE 9. INTAKES OF CALCIUM, PHOSPHORUS AND NITROGEN PREDICTED FOR EQUILIBRIUM; AGING WOMEN; SUCCESSIVE DECADES

AGE RANGE	NO.	INTAKE PREDICTED FOR EQUILIBRIUM		
		Ca.	P	N
yrs.		*gm/day*	*gm/day*	*gm/day*
30–39	25	0.88	1.25	10.64
40–49	34	0.87	1.32	10.98
50–59	40	0.83	1.42	11.20
60–69	25	0.92	1.51	10.76
70–79	11	0.73	1.13	9.42

calcium and phosphorus which are seen in the controlled experiment (16–18).

Of considerable interest in this experiment are the differences in both food intake and economy of utilization seen in the various age groups. Calories, protein and phosphorus intakes are reduced in each successive decade. The intake of each nutrient studied was highly correlated with the intake of calories. The fact that more than a third of the subjects were overweight might suggest that caloric intakes, at least, are higher than required.

No appreciable reduction in calcium intakes occurred until the eighth decade, and predicted requirements for mean equilibrium do not suggest that calcium intakes should be reduced until that time. However, individual women were in equilibrium with respect to all three nutrients over a wide range of intakes when a minimum of 1500 calories were ingested. Women under 40 and over 70 appeared to be able to adjust metabolic demands and intake with a

degree of individual variation in keeping with the findings of other laboratories (16–18) when allowances are made for differences in experimental pattern between a self-selected and a controlled diet study. However, individual variation increased after the age of 40, was marked in the sixth decade and decreased somewhat in the seventh. Certain subjects maintained economy in utilization of nutrients but the number of those who did not increased to a peak in the sixth decade and the degree of inefficiency also increased as measured by greater losses of each of the three nutrients at a given intake. This generalization is further supported by the apparently higher requirements for mean equilibrium of nitrogen and phosphorus which were found for these age groups. Greater variability in response, particularly in the direction of poorer use of nutrients, might be an index of higher requirements of a population sample (table 9). The adult who has been previously well nourished can apparently accept a considerable period of nutrient losses before overt symptoms occur and there is no easy clinical way to identify these inefficient individuals until frank deficiency symptoms are seen. At this time body stores are exhausted, and since depletion has occurred over a long period of time, repletion and recovery of well being can be slow and discouraging. There is no measure of the permanent damage done or how much the aging process may be accelerated in an individual who is in the middle age ranges at the time that depletion occurs. From this point of view, a realistic set of dietary standards in a civilization with a generous food supply must be high enough to include the members of a population who, either because of previous depletion, or inefficiency, require larger amounts of specific nutrients to maintain body structure and function.

The greater variability in the use of nutrients by certain individuals of the 40–59 age groups is understandable in relation to the glandular changes which are characteristic of the menopause and the variability in symptoms among women. The persistence of variability into the next decade may measure a failure of adjustment following the menopause but it has been pointed out (12) that these ages bring both social and economic adjustments associated with retirement and curtailed income, death of spouse and occasionally the breaking up of the home with the older woman the sole survivor of her generation. If she survives until age 70, it is with a more stable adjustment to food and its use; or perhaps she survives because she is one of the group who have been able to adjust to change throughout her life.

The findings of this study would suggest that the present recommended allowances for calcium and nitrogen are not excessive for aging adult women under the conditions of self chosen diet described in this paper. Reduced caloric intakes without attention to selection of nutrients may be the key to understanding many cases of nutritional instability. It also is suggested that the period 40–70 represents a time of metabolic stress for many adult women and that dietary instruction must be as individualistic as medical care if the nutritional reserves of the woman are to be protected.

REFERENCES

1. Carlson, A. J. and F. Hoelzel. *J. Nutrition* 34: 81, 1947.
2. McCay, Clive M., G. Sperling and L. L. Barnes. *Arch. Biochem.* 2: 469, 1943.
3. Wood, T. A. as quoted by Sherman, H. C. *Chemistry of Food and Nutrition*, 7th ed., New York: Macmillan, 1946, p. 512.
4. Stearns, G. *J. Lab. & Clin. Med.* 14: 954, 1929.
5. Fiske, C. H. and Y. Subbarow. *J. Biol. Chem.* 66: 375, 1925.
6. *Composition of Foods*. Bureau of Human Nutrition and Home Economics, Handbook 8. Washington, D. C.: USDA, 1950.
7. Bowes, A. de P. and C. F. Church. *Food Values of Portions Commonly Used*. 6th ed. Philadelphia: Campus Offiset Press, 1946.
8. Leichsenring, J. M. and E. D. Wilson. *J. Am. Dietet. Assoc.* 27: 386, 1951.
9. Knapp, E. L. *J. Clin. Investigation* 26: 182, 1947.
10. Ohlson, M. A., L. Jackson, R. Marin, D. Dunsing and W. D. Brewer. *J. Gerontol.* 6: 131, 1951.
11. Roberts, P. H., C. H. Kerr and M. A. Ohlson. *J. Am. Dietet. Assoc.* 24: 292, 1948.
12. Ohlson, M. A., L. Jackson, J. Boek, D. C. Cederquist, W. D. Brewer and E. G. Brown. *Am. J. Public Health* 40: 1101, 1950.
13. Kountz, W. B. *J. Gerontol.* 6: 114, 1951.
14. Leverton, R. M. and M. R. Gram. *J. Nutrition* 39: 57, 1949.
15. Swanson, P. P. and H. E. Clark. *Ann. Rev. Biochem.* 19: 235, 1950.
16. Bricker, M., H. H. Mitchell and G. M. Kinsman. *J. Nutrition* 30: 269, 1945.
17. Hegsted, D. M., A. G. Tsongas, D. B. Abbott and F. J. Stare. *J. Lab. & Clin. Med.* 31: 1, 1946.
18. Steggerda, F. R. and H. H. Mitchell. *J. Nutrition* 17: 253, 1939.

CHANGES OF BODY COMPOSITION IN MAN DURING MATURITY AND THEIR NUTRITIONAL IMPLICATIONS[1]

JOSEF BROŽEK

From the Laboratory of Physiological Hygiene, University of Minnesota, Minneapolis

IN MAN, maturity covers a span of some 30 or 40 years (roughly from 20 to 55 years of age). The term suggests that this period represents a sort of 'steady state' in the individual's biological history. This illusion is likely to be due to two sets of factors. Firstly, maturity is the longest of man's life periods and the changes proceed at slower rates than during childhood and adolescence on the one end, and senescence on the other end of the candle of life. Secondly, the period has been studied much less intensively and systematically than the initial or the final phase of man's ontogenesis. Nevertheless, the fact remains that there is no 'steady state.' Life moves on, the biological clock keeps on ticking. It is only when we compare at brief intervals the individual's status, including his body composition, that the changes appear imperceptible.

The general age trend of relative fatness in man is known, with the early childhood and late maturity as peaks in the fat content of the body (1). However, detailed quantitative data covering the whole life span are non-existent. As is the case of most areas of human biology, up to now there has been little information on the measures of fatness beyond the age of college students.

It has been only during the last decade that methods for a fairly satisfactory analysis of body composition *in vivo* have become available (2, 3). The separation of the fat and fat-free component of body weight, based on the principle of specific gravity, developed by Behnke and his co-workers (4–9) opened an important avenue to metabolic, physiological, and nutritional studies. This new tool was supplemented and further extended by taking into account total body water (10–13) and the partial 'fluid spaces' (14–18). These developments are of particular significance to the science of nutrition. In a way, this completes a full turn on the helix of the history (19) of nutritional research. Following the early phase concerned with calories, the emphasis was shifted successively to proteins, minerals, vitamins and amino acids. The new technical developments for the analysis of body composition bring again into the foreground the matter of calories.

It has been stated that in this country the excessive caloric intake, resulting in 'overnutriture' (relative obesity), is the most widespread and most devastating nutritional disorder. The toll of the so called 'degenerative diseases,' in the etiology and course of which obesity seems to play a definite though not well understood role (20, p. 193), is becoming steadily heavier. Retardation of the time of onset of clinical symptoms and reduction of the severity of arteriosclerosis and hypertension, of cancer, and of diabetes constitute a major challenge to medical science. In this task the control of obesity is likely to be of central importance (20a).

While 'overweight'—defined as a positive departure from the weight 'standard' for height, age and sex—represents a useful index of obesity for actuarial studies, physiological analysis of the interrelationships between gross tissue morphology and the functional characteristics of the aging organism requires a more precise description of calorie nutriture. It will be agreed readily that composition of the body mass constitutes a basic criterion of the nutritional status of the organism with reference to calories. The separation of the body weight into the fat component and the lean body mass represents the first and fundamental step in the quantitative description of the compartments of the body.

During maturity there are not only a continued increase in gross body weight but also changes in the relative amounts of different tissues. It has been shown in a preliminary communication (21) that middle-aged men, retaining a central position in the distribution of weights for individuals of a given height, differ markedly in their physique from men in their early twenties. The two groups studied consisted of men 19–26 and 45–55 years of age, drawn from larger samples of college students and business and professional

[1] The work reported here was supported by a research grant from the National Institutes of Health, Public Health Service.

men by selecting those individuals whose actual body weight was within ±5 per cent of the standard weight for height and age. The mean values of the relative body weight, obtained by expressing the actual weight as percentage of the standard weight, were very close in the two age groups, with values of 100.2 and 100.0, respectively. Nevertheless, the men's fatness—evaluated in terms of the thickness of skinfolds and estimated from the specific gravity—was strikingly greater in the older men. It was pointed out that in aging individuals the upward trend in the 'normal' body weights, defined actuarially as the mean value for groups of men of a given height and age, obscures the profound changes in the relative value of body compartments.

In the second preliminary study (22) a group of younger men (mean age 21.8 years) was matched individually by one or two older men (mean age 50.3 years) of the same height *and* same absolute body weight. There still remained marked differences in fatness, reflected in specific gravity, skinfolds, and even such indirect criteria of fatness (23) as the difference between the chest and abdomen circumference.

As far as the principal criterion of leanness-fatness, i.e. the specific gravity, is concerned, the data on which both of these studies were based suffered from the fact that residual air volumes, needed as a correction applied to the underwater weight in computing the body volume, were not determined in each individual case; instead, an average correction was used. This is particularly important as the residual air itself exhibits an age trend (24).

In the last two years data have been accumulated which do not suffer from these limitations. In the present paper the matter of changes in body composition with age, with emphasis on the separation of the changes in the fat and the lean body mass, is submitted to a comprehensive examination.

METHODS

There is a variety of techniques which are of potential or demonstrated value as measures of fatness (2). In the present study we have relied principally on determinations of specific gravity (4) and on measurements of skinfolds (25). In computing the density of the whole body, the body volume is determined on the basis of the Archimedean principle by weighing the subject under water. The details of underwater weighing and of determining the residual air volume at the time of submersion, using the nitrogen dilution method, were presented elsewhere (26).

The thickness of skinfolds was measured by a caliper with flat, circular contact points (diameter of 7.0 mm., area of 38.5 mm²). The skin is compressed between the thumb and the index finger of the left hand, with the 'bite' covering a distance of about 8 cm. The thickness of the lifted fold represents double the value of the skin plus the subcutaneous tissue. The caliper is activated by a coil spring. The coil consists of many rings in order to minimize the increase in tension due to the compression of the spring as the distance between the contact points of the caliper increases.[2] In the present model, the initial tension, i.e. tension necessary for the minimal opening of the jaws of the caliper, was 208 gm. (5.4 gm/mm²). A further increase of the distance was associated with a small, fairly linear, rise in the spring tension. At the opening size of 5.5 cm. which is close to the upper end of the range of the skinfold thickness values, the tension increased to 270 gm. (corresponding to about 7.0 gm/mm.²). This increase results in a negligible distortion of the thickness values.

The skinfold measurements were made at 7 points: *1*) abdomen—to the right of the navel; *2*) chest—above and to the right of the right nipple; *3*) back—immediately below the right scapula; *4*) arm—on the back side, half-way up the upper arm; *5*) thigh—above the knee cap; *6*) midline—above the iliac crest, along the mid-axillary line; and *7*) chin—below the chin, along the anterio-posterior line. All skinfolds were vertical, with the thickness measured in the horizontal direction.

SUBJECTS

The subjects used in the present study fall into two age groups. The younger men were students at the University of Minnesota at the time of the first examination, with an age range of 18–25 years. A good proportion of them were veterans. The overwhelming majority of the men were single. The older men, with an age range of 45–54 years at the start of the study, were drawn from the ranks of business and professions in the Twin Cities.

These men volunteered for a long-term, inter-

[2] The Committee on Nutritional Anthropometry (Food and Nutrition Board, National Research Council) is taking steps toward a design and commercial manufacture of a standard caliper with a constant tension throughout the whole range of thickness measurements.

disciplinary study on aging, focussed on the development of 'degenerative' diseases of the cardiovascular system (27). All applicants were carefully screened and the individuals incorporated in the present study may be considered, in a rigorous sense, as being clinically normal.

RESULTS

The basic data are presented in tables 1 and 2. In all 5 samples the mean relative weight closely approximates 100. That is, the actual mean weight is very close to the 'standard' weight for the particular age and height in the three groups. criteria, specific gravity and thickness of skinfolds, indicate marked differences between the younger (23–29) and the older (48–57) individuals, the latter exhibiting a striking increase in fatness. The specific gravity and the derived fat content showed consistent and continuous change within the age group 48–57. Most of the skinfolds differentiated well between the younger and older individuals but the differences between the two groups of older men were not consistent. The skinfold measured above the knee cap, indicated in the table as 'thigh,' shows only a questionable increase with age. It was somewhat

TABLE 1. AGE DIFFERENCES IN FATNESS, SAMPLES WITH AN AVERAGE STANDARD WEIGHT (RELATIVE WEIGHT ≐ 100)

AGE GROUPS	23–29		48–52		53–57	
N	25		44		34	
	M	*SD*	*M*	*SD*	*M*	*SD*
Age	25.2	1.9	50.0	1.5	54.6	1.2
Height	176.5	5.8	175.4	5.7	174.8	6.2
Gross body wt., kg	70.6	8.3	75.9	6.7	76.0	7.5
Relative wt. (gross as % of standard)	99.9	9.7	99.9	5.1	100.2	5.1
Specific gravity[1]	1.0695	0.0127	1.0500	0.0115	1.0475	0.0114
% fat	14.4	6.2	24.0	5.7	25.2	5.7
Fat, kg	10.2	5.1	18.2	5.1	19.2	5.6
Fat-free wt., kg	60.4	5.1	57.7	5.4	56.8	5.3
Skinfolds, mm.						
Abdomen	19.1	7.8	26.0	5.6	26.0	6.6
Chest	16.3	7.1	26.9	6.8	27.9	7.4
Back	14.6	5.3	20.4	5.4	21.5	7.1
Arm	13.7	3.0	16.3	4.2	15.4	4.0
Thigh	8.9	1.9	9.5	2.4	9.4	2.5
Midline	17.1	2.3	20.0	5.6	19.6	5.3
Chin	8.9	1.9	10.8	3.5	11.0	2.7

[1] The value of residual air, used for correcting the underwater weight, has been determined for each individual at the time of submersion.

In the age groups 48–52 and 53–57 (table 1), the sample was restricted to individuals within ±10 per cent of the standard weight. In the younger group (23–29) both lighter and heavier subjects were included, and this is reflected in the larger standard deviation of the relative weight in this group. It should be noted that determinations of residual air were carried out in all individuals represented in table 1 and that these values were obtained for conditions existing at the end of maximum expiration at which time the underwater weight was read on the scale.

The mean values of both types of fatness of a surprise to find that the skinfold measured above the iliac crest and under the chin were also relatively insensitive indicators of changes in total fatness with age. The skinfold measured at the chest, above nipple, showed the largest increment (11.6 mm.). This is in keeping with the fact that the chest skinfold showed the highest correlation with specific gravity, with $r = -.857$ in a group of younger and $r = -.682$ in a group of older men (25).

Specific gravity and the derived fatness values for age groups 18–22 and 45–47 are presented in table 2. For these men individual determinations

of residual air were made at a later date; the correction was applied to an earlier set of data on gross body weight and on weight under water at the time of maximal expiration. The means indicate a difference of 12.3 per cent of fat between the two age groups. The mean values are lower than those obtained for the next higher age groups and fit well in the general pattern of table 1.

In the two groups described in table 2, the skinfolds were measured at a smaller number of locations—3 in the age group 18–22, 5 in the age group 45–47. The differences between the age groups were of the same magnitude as in table 1 (+9.1 mm. for abdomen, 10.3 mm. for chest, 7.5 for back) but the absolute values cannot be readily compared with those in table 1

TABLE 2. AGE DIFFERENCES IN FATNESS. SAMPLES WITH AN AVERAGE STANDARD WEIGHT (RELATIVE WEIGHT ≐ 100)

AGE RANGE	18-22		45-47	
N	21		29	
	M	*SD*	*M*	*SD*
Age	20.3	1.6	46.0	0.8
Height, cm.	177.0	6.3	176.1	5.5
Gross body wt., kg.	68.7	9.1	76.6	7.9
Relative wt.	100.0	10.7	100.2	6.2
Specific gravity[1]	1.0789	0.0123	1.0535	0.0101
% fat	9.9	5.9	22.2	5.1
Fat, kg.	6.8	4.8	17.1	4.8
Fat-free wt., kg.	61.9	6.1	59.5	2.1

[1] The value of residual air, obtained for each individual during submersion in 1951, was applied to the earlier data (1947 for young men, 1948 for the older men) of gross underwater weight.

because a caliper has been used which differed in the size of the contact points and the strength of the spring (25).

The mean values of body fat, expressed as percentages of gross body weight (tables 1 and 2) are plotted in figure 1. A least-square line, fitted to these points, is defined by *equation 1:*

$$\hat{Y} = 2.3692 + 0.4276\,X \qquad (1)$$

where X = age, in years and $\hat{Y}$ = fat percentage predicted from age. However, it is not likely that the rate of change in fatness would be uniform and the trend is better described by a parabola defined by *equation 2:*

$$\hat{Y} = -5.55564 + 0.92836\,X - 0.006776\,X^2 \qquad (2)$$

This curve is presented in figure 1.

The 5 samples (tables 1 and 2) vary somewhat in the mean height. Consequently, the body weights—and their fat and fat-free fractions—cannot be directly compared. In part, the lower height of the older men reflects a true decrement of stature with age, in part, the differences may be due to sampling. In order to obtain a set of directly comparable body weights, we have tabulated the standard weights for individuals of the same height (176 cm.) at age levels 20–55 (table 3, first column). The 'standard' fat percentages were obtained from *equation 2.* The third and

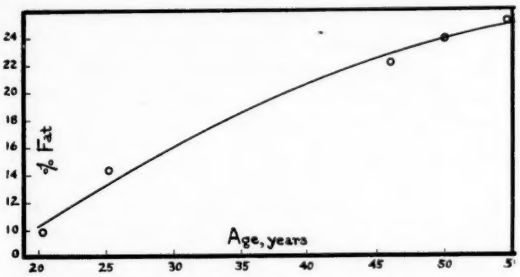


Fig. 1. MEAN VALUES of body fat, as percentage of body weight, during maturity. The subjects are clinically normal men.

TABLE 3. AVERAGE ESTIMATED CHANGES IN BODY COMPOSITION DURING MATURITY (20–55 YEARS). STANDARD WEIGHTS REFER TO MEN 176 CM. TALL.

AGE	STANDARD WT.	STANDARD FAT	FAT	FAT-FREE WT.
	kg.	%	*kg.*	*kg.*
20	67.6	10.30	6.96	60.6
25	69.9	13.42	9.38	60.5
30	71.3	16.20	11.55	59.8
35	72.9	18.64	13.59	59.3
40	74.1	20.74	15.37	58.7
45	75.5	22.50	16.99	58.5
50	76.0	23.92	18.18	57.8
55	76.8	25.01	19.20	57.6

fourth column were derived from the data in columns 1 and 2 of table 3.

The increase of body weight, during the time when stature has reached its maximum ('adult') level and then actually begins to decline, is a well established fact, observed in numerous studies. However, until recently no adequate data have been available for the analysis of these weight increments. The present study confirms the findings (22) that the average weight increments are due solely to the increase in body fat. In fact, the increase of the fat fraction—on the absolute (as well as on the relative basis—is

greater than the increment in gross body weight. This is reflected in the actual decrement of the lean ('fat free') body mass.

We have complete confidence in the validity of the fat percentages in the age groups 23–29, 48–52, and 52–57 (table 1) in which individual values of the residual air were determined at the time of the underwater weighing. As we have noted, in the age groups 18–22 and 45–47 (table 2) the residual volumes were determined at a later date and the values were applied to the earlier figures of underwater weight. The mean value of the residual air in the group of 21 men, 18–22 years old at the time of underwater weighing, equaled 1.464 liters, with a standard deviation of the series equal to 0.328 liters. It is legitimate then, to compare the specific gravity of this smaller sample with a larger sample of men of the same age for whom the specific gravity was obtained by using an average correction (1.5 l.) as the value of the residual air. In a group of 48 men, with average relative weight of 100.0 and a range from 95.0–104.9, the specific gravity was almost identical (table 4, first column) with that obtained in the 21-man sample (table 2).

TABLE 4. AGE DIFFERENCES IN FATNESS. SAMPLES WITH AN AVERAGE STANDARD WEIGHT (RELATIVE WEIGHT = 100)

AGE GROUPS	18–23		20–33	
N	48[1]		20[2]	
	M	*SD*	*M*	*SD*
Age	20.7	1.8	25.4	3.9
Height	178.0	6.7	177.6	6.0
Gross body wt., kg.	69.5	5.5	71.5	5.3
Relative body wt.	100.0	2.7	100.2	5.4
Specific gravity	1.0791	0.0119	1.0686	0.0114
% fat	9.7	5.6	14.8	5.5
Fat, kg.	6.7	4.1	10.6	4.6
Fat-free wt., kg.	62.8	5.3	60.9	3.6

[1] University students. A constant value of 1,500 cc. was used as correction for residual air. [2] Conscientious objectors. A constant value of 1,450 cc. was used as correction for residual air.

For the mid-twenties we have available an independent check on the average value of specific gravity. Again, the sample consists of individuals with the average body weight equal to the standard weight (for height and age). This set of data was obtained during the control period of a starvation-rehabilitation experiment (28). The subjects were normal young men, of much the same socio-economic background as our samples of 23–29 year old subjects. The mean actual fat content was 14.8 per cent (table 4, second column)—a value somewhat higher than the predicted value of 13.7 and close to the actual mean fat percentage of the sample of the 23–29 year old men (14.4%).

For a group of 81 men, with an average age of 26.6 years, Osserman *et al.* (29) reported specific gravity of 1.068, yielding body fat of 15.0 per cent. The mean height was 176.6 cm. (30), actual weight 75.3 kg., standard weight 71.0 kg.; consequently, the men were somewhat overweight (mean relative weight of 106.0). The predicted fat percentage for individuals of the same age and of standard ('normal') body weight is 14.3 per cent. This and the data presented in table 4 indicate that the general magnitude of the predicted fat percentages is not far out of line with the empirical checks.

The comparisons of different age groups made in tables 1–4 were made on the basis of relative body weight. This is proper when we wish to trace the average trends. It is for this reason that the samples were selected in such a way that the actual body weight coincided with the standard weight for age and height.

Of some interest is a comparison of body composition in two age groups matched on the basis of absolute body dimensions (height and gross body weight). The matching of individual cases was not always perfect. However, the group averages of the two variables used as criteria for matching are very close to each other. The data are presented in table 5.

Even under these conditions, involving a comparison of overweight younger men (mean relative body weight = 106.2) and somewhat underweight older men (mean relative body weight = 97.2), the latter group shows a definitely greater fatness.

ETIOLOGY OF THE AGE CHANGES

The presentation so far has been concerned with the description of the age trends in body composition. Logically, the next task is an analysis of the factors responsible for the observed changes. It should be pointed out at the outset that the present possibilities of interpreting these changes and, in particular, of assigning numerical values to the separate factors, are severely limited. Important data needed for a valid and comprehensive analysis are lacking.

To start with, we do not know with any precision the caloric intake of the subjects and there are no data in the literature which would allow one to make even an intelligent guess as to the age differences in caloric consumption. To make matters worse, the information on the caloric content of the diet would make little sense physiologically in the absence of information about the caloric equivalent of the individuals' activity. That the activity level is changing with age is known from common-sense observation. differences in the caloric equivalents of physical activity.

It may be of interest to note that in experimental semistarvation, which in a variety of ways resembles aging, a much larger decrement took place in the energetic cost of physical activity than in the basal metabolism component. The starvation value of the caloric equivalent of 'physical activity' represented only 30 per cent of the control value (32). It is very likely that the decrement with age will again be considerably

TABLE 5. YOUNGER AND OLDER MEN MATCHED ON BASIS OF HEIGHT AND GROSS BODY WEIGHT

AGE GROUPS	22-29		48-57	
N	33		33	
	M	*SD*	*M*	*SD*
Age	25.2	2.0	52.1	2.7
Height, cm.	177.1	6.0	176.9	6.0
Gross body wt., kg.	75.6	11.8	75.7	11.4
Relative body wt.	106.2	10.6	97.2	12.8
Specific gravity[1]	1.0651	0.0152	1.0528	0.0119
% fat	16.5	7.5	22.6	5.9
Fat, kg.	12.5	7.4	17.1	6.7
Fat-free wt., kg.	63.1	6.8	58.6	7.0
Skinfolds, mm.				
Abdomen	22.6	10.7	24.3	9.0
Chest	19.3	9.7	24.5	8.9
Back	18.3	9.3	18.5	7.7
Arm	15.4	5.7	14.8	5.6
Thigh	10.2	3.3	8.8	3.2
Midline	20.7	9.8	18.3	7.2
Chin	10.1	4.1	10.6	4.2

[1] Residual air determined individually at the time of submersion.

This fact was also brought out in systematic studies of behavior changes with age (31). Thus, when the differences between our two age groups consisting of college students and middle-aged business and professional men were studied by means of personality questionnaires, 30 per cent of the older men but only 7 per cent of the younger group noted that they "would rather sit inside and read, listen to the radio or talk than to take walks or play active games" (critical ratio of the difference = 4.78). The item "I know nothing I'd rather do than spend a quiet evening at home" was answered "Yes" by 62 and 23 per cent of the older and the younger men, respectively (C.R. = 6.66).

These data clearly indicate the age trends but do not allow for even a gross estimate of the more marked in this characteristic than in basal metabolism.

The basal metabolism values for the two age groups, matched for height and gross body weight (see table 5), are given in table 6. The basal caloric expenditure per 24 hours corresponds to 1774 Cal. and 1516 Cal., respectively. In the Mayo standards for basal metabolism, given as calories per square meter of body surface per hour, the age difference corresponds to 9.1 per cent of the value for the younger men. In the present series the difference in the basal caloric output represents 14.5 per cent.

Nothing would be gained by expressing the caloric expenditure in terms of the body surface as the latter is estimated from height and weight, which are identical in the two groups. Physi-

ologically, 'active tissue mass' (obtained as body weight minus weight of bone minerals, extracellular fluid, and fat) appears to provide the optimal reference value for comparisons of basal caloric outputs. Unfortunately, two of the components—minerals and, more importantly, the magnitude of the fluid compartment—are not known.

Even if these values were available, some 'error of measurement' would be introduced by the fact that the composition of the lean body mass itself is subject to age changes. This affects the very basis on which estimations of excess fat from specific gravity are founded (9). However, it appears that these changes are relatively small and may include processes which are compensatory, as was the case in semistarvation (28, p. 283). The figures presented in this paper may require correction but the basic picture and conclusions are not likely to be altered.

TABLE 6. AGE DIFFERENCES IN MEAN BASAL OXYGEN CONSUMPTION IN TWO AGE GROUPS OF SAME GROSS BODY SIZE (see table 5).

AGE RANGE.....	22-29	48-57	$\overline{\Delta}$	$\overline{\Delta}$%
cc/min.....	254	217	37	14.6
cc/min/kg. gross body wt.....	3.36	2.87	0.49	14.6
cc/min/kg. fat-free body wt..	4.02	3.70	0.32	8.0
cc/min/kg. 'active tissue'...	5.98	5.73	0.25	4.2

Aware of the approximative nature of the calculations, one may attempt to estimate the active tissue mass in the younger men. The bone minerals were calculated as 4 per cent of the standard body weight (70.8 kg.), i.e. about 2.8 kg. The extracellular fluid, as thiocyanate space, was shown to correspond under normal conditions of hydration to about 23.6 per cent of the body weight, i.e. 17.8 kg. The fat amounted to 12.5 kg. This yields an active tissue mass of (75.6 − 2.8 − 17.8 − 12.5) = 42.5 kg. For the older individuals the respective values are (75.7 − 2.8 − 17.9 − 17.1) = 37.9. Expressing the basal O_2 consumption in reference to the values of active tissue mass, the difference between the two age groups is markedly reduced (table 6).

Nevertheless, the difference is not eliminated. This is an indication that aging is associated with a real slowing down of the metabolic rate in the tissues. It should be pointed out that the change of the metabolic rate in the 'active' tissues is small during the period of maturity.

In the etiology of the observed decrease in the active tissue mass during the process of growing older, the decreased physical activity is likely to be the principal factor. The 'disuse atrophy' accounts for some and most likely for all of the loss of active tissues inferred from our studies on young and middle-aged men, although hormonal effects cannot be ruled out. Our sample was drawn from 'middle-class,' urban dwellers. We should like to have the study repeated on a sample of manual laborers (in a literal sense, not on 'ditch diggers' who operate steam shovels or on 'hod carriers' who pour ready-mixed cement from a truck into a mold). If disuse atrophy is the principal factor, there should be definite differences in the lean body mass (and the 'active tissues') between groups of individuals engaged in varying amounts of physical work.

This question was examined by making a comparison between a group of physically active and inactive individuals drawn from our larger sample of normal middle-aged business and professional men. The physically 'active' group continued during adulthood to participate in recreational activities involving a fair amount of physical exertion while the 'inactive' group had always been relatively sedentary. Strangely enough, comparisons made with reference to 56 separate characteristics revealed only negligible differences in physiological functions (34). At the same time there was a definite difference in the size and, more importantly perhaps, in the composition of the body. The data, calculated on the basis of recent measurements, including direct and individual determinations of the residual air, are presented in table 7.

The active individuals are heavier and appear 'overweight.' However, and this is a significant point, noted earlier by Welham and Behnke for professional football players (7), the excess weight is due to the greater fat-free body weight, not to an excess of fat. The mean specific gravity is higher than the mean of the inactive individuals; consequently, the estimated fat content—on a relative as well as on the absolute basis—is lower. The skinfolds are larger in the inactive groups but the interindividual variability is large and the single differences are not significant statistically. The difference in the specific gravity of the two groups approaches the 5 per cent level significance of the F-test ($F = 3.93$, $F_{0.05} = 4.00$). The difference in lean body mass exceeds

the 1 per cent level of statistical significance ($F = 10.01$, $F_{0.01} = 7.08$). These data confirm the hypothesis that the decrement in fat free body weight (see table 3) represents essentially partial 'disuse atrophy.'

The etiology of the average age trend toward an increase of the fat depots (and of gross body weight) is simple. It reflects a small but definite excess of caloric intake over the caloric output present over a period of years. On the 'down grade' we are less sure. It is generally believed that in old age there is a gradual loss of body fat.

COMMENT

In accordance with the spirit of the times the unitary concept of 'nutritional status,' corresponding to a global and qualitative evaluation, is giving way to a detailed quantitative description. The present study, carried out with reference to the changes in body composition taking place during the process of aging, may serve to illustrate the trend toward increased precision in the description of the nutritional status (nutriture), specifically of its caloric aspect.

TABLE 7. DIFFERENCES IN BODY COMPOSITION OF PHYSICALLY 'ACTIVE' AND 'INACTIVE' MIDDLE-AGED MEN

GROUP	ACTIVE		INACTIVE	
N	29		27	
	M	*SD*	*M*	*SD*
Age	52.9	3.2	52.2	2.9
Height, cm	176.1	6.3	176.2	5.0
Gross body wt., kg	81.5	9.2	78.3	12.3
Relative body wt. (actual as % standard)	106.2	11.9	102.2	13.6
Specific gravity[1]	1.0493	0.010	1.0435	0.012
% fat	24.3	5.2	27.3	6.3
Fat, kg	19.8	5.8	21.4	7.8
Fat-free wt., kg	61.7	7.0	56.9	5.7
Skinfolds, mm.				
Abdomen	26.8	4.9	28.2	9.0
Chest	28.3	5.6	29.3	8.8
Back	21.2	5.3	23.0	10.0
Arm	16.4	3.3	16.5	5.6
Thigh	10.4	2.9	11.0	4.1
Midline	20.7	4.5	20.9	9.0
Chin	11.0	2.4	12.5	5.4

[1] Residual air determined individually at the time of submersion.

This process, not adequately documented in quantitative terms, seems to involve both the deep and the superficial fat depots (35, p. 158). It is not clear what factors are involved in the reversal of the fat gain observed during maturity. There is no evidence for the idea that this represents a compensation for an inadequate intake of calories in general or of fat in particular, but the whole matter requires further factual study. Of opinions, frequently contradictory, there is no dearth in this or any other area of gerontology. The questions of taste ('appetite') for fat, presumably increased in older people; of actual intake of fat; and of caloric intake must be handled separately.

Its principal factual contribution is a quantitative documentation of the fact that the increment of body weight with age is a complex resultant of simultaneous changes in several body compartments. These processes are going on not only at different rates but even in different directions. We have established, as a first approximation, that the rise in body weight (d_W) during maturity is composed of an increase in body fat (d_F) and a decrease in the lean body mass (d_L)

$$d_W = d_F + d_L \qquad (3)$$

Over the period from 25–50 years the d values, derived from table 3, are as follows:

$$6.1 = 8.8 + (-2.7) \qquad (4)$$

The lean body mass is a morphologically complex variable, with bones, extracellular fluid, and the 'active tissues' (nerves, glands, muscles) as its compartments. The present data do not allow a precise partitioning of the decrement in the overall fat-free body weight into these components. However, it is likely that a decrease in the muscle mass, replaced by fatty tissues, accounts for most of the change. Information on age changes in body fluids is needed to add precision to the present picture of the trends in body composition. For this purpose data on both the total body water and the separate fluid spaces are necessary. The separation of extracellular and of intracellular water is particularly interesting, and provides an opportunity for a check and cross validation of the concept and the numerical value of 'active tissue' mass.

It is one of the items of 'general information,' difficult to trace in its origin (cf. 36) that man (and other mammals) shows a steady decline in the proportion of total water in the body as he grows older. During maturity this is due largely to the increasing fat depots which contain a much smaller amount of water (6–20%) than the muscle (75% water); (37). In the calculation carried out for comparison of the basal metabolic rate in the older and younger men we have postulated that the extracellular fluid space has not changed during maturity. Data on this point as well as on total body water are being accumulated at least in one research center but the impression that the extracellular space remains essentially unaltered is considered as tentative.

The data on age trends in body composition are limited in several other ways. For one thing very little is known about the changes in body composition after the age of 60. The actuarial height-weight tables in common use stop at the age of 55 years and one is uncertain even of changes in gross body weight during old age. The present study is confined to events taking place during the 40-odd years of maturity. Even in this range data are lacking for the middle part of this period.

Furthermore, we can say next to nothing about age trends in the body composition of women during and beyond maturity. All systematic and quantitative observations were made so far on men. The available information indicates that females have a larger amount of fat than males at all age levels (28, p. 173). Actuarial tables indicate that women have the tendency to gain weight during maturity, the average gains being of similar magnitude as in men; when expressed in reference to body weight, the gains are larger. This may be interpreted as an indication of a still steeper increase in the fat content than has been observed in men.

Nutritionally, the principal conclusion to be drawn from the data presented in the body of this paper is that the average caloric intake during maturity tends to be larger than the actual caloric need. The depot fat, although having definite other functions in the organism (such as insulation), serves primarily as a fuel reserve. Except under catastrophic conditions resulting in famine, the accumulation of fat can hardly be regarded as biologically desirable even though the increased fat content of the body is a 'normal' process in the statistical sense.

The widely used dietary allowances, recommended by the Food and Nutrition Board of the National Research Council (38) take into account sex, including the state of pregnancy and lactation and the activity level of the adult individuals, but do not provide an adjustment for age. The FAO Committee on Food Requirements (39), concerned with population rather than individual caloric needs, felt that the age factor cannot be neglected and that a scaling of caloric requirements of adults according to age is necessary. In view of the very large differences in the age composition of different populations and in view of the established decrement in calorie needs during maturity such a consideration appears fully justified. The evidence is not cited in detail but several points are brought out. The decrement in basal metabolism with age is substantially linear; at the age of 65 years it is about 20 per cent less than at the age of 25 (15 per cent according to the Mayo standards). A tendency for activity to diminish with increasing age is noted; the decline in the activity component of the total energy expenditure is relatively larger than the decrement in basal metabolism. The trends of food consumption have not been charted satisfactorily but the available surveys of the food consumption of elderly people indicate a smaller caloric intake than in the younger adults.

The FAO Committee was aware of the lack of precise data which would serve as a firm basis for making an age correction in the tables of caloric requirements. It was suggested to reduce

the tabular values by 7.5 per cent of the requirement at age 25 for each 10 years beyond the age of 25. On this basis the average requirement at 35, 45, 55, and 65 years would be reduced to 92.5, 85.0, 77.5 and 70.0 per cent of the value for the 25 year old invididuals. No magic or infallible quality is assigned to this correction. Expressed in numerical terms, it is amenable to adjustment in the light of further research.

The study of body composition and the data on which the present report rests represent one facet of a cooperative research involving, in the ranks of the senior staff, Drs. H. L. Taylor, Joseph Anderson, Ernst Simonson, Carleton B. Chapman, and Mrs. Nedra Foster. Dr. Ancel Keys, Director of the Laboratory, was associated particularly closely with this phase of the overall research program. Mr. Walter Carlson assisted with the specific gravity determinations.

REFERENCES

1. Berg, W. Z. *Morphol. Anthropol.* 13: 305, 1951.
2. Brožek, J. and A. Keys. *Nutr. Abstr. Rev.* 20: 247, 1950/51.
3. McCance, R. A. and E. M. Widdowson. *Brit. Med. Bull.* 7: 297, 1951.
4. Behnke, A. R. *Harvey Lectures* 37: 198, 1941/42.
5. Behnke, A. R. *Medicine* 24: 359, 1945.
6. Behnke, A. R., B. G. Feen and W. C. Welham. *J. A. M. A.* 118: 495, 1942.
7. Welham, W. C. and A. R. Behnke. *J. A. M. A.* 118: 498, 1942.
8. Rathbun, E. N. and N. Pace. *J. Biol. Chem.* 158: 667, 1945.
9. Morales, F. M., E. N. Rathbun, R. E. Smith and N. Pace. *J. Biol. Chem.* 158: 677, 1945.
10. Pace, N. and E. N. Rathbun. *J. Biol. Chem.* 158: 685, 1945.
11. Pace, N., L. Kline, H. K. Schachman and M. Harfenist. *J. Biol. Chem.* 168: 459, 1947.
12. Messinger, W. J. and J. M. Steele. *Proc. Soc. Exper. Biol. Med.* 70: 316, 1949.
13. Steele, J. M., E. Y. Berger, M. F. Dunning and B. B. Brodie. *Am. J. Physiol.* 162: 313, 1950.
14. Henschel, A., O. Mickelsen, H. L. Taylor and A. Keys. *Am. J. Physiol.* 150: 170, 1947.
15. Berger, E. Y., M. F. Dunning, M. F. Steele, R. Jackenthal and B. B. Brodie. *Am. J. Physiol.* 162: 318, 1950.
16. Levitt, M. R. and M. Gaudino. *Am. J. Med.* 9: 208, 1950.
17. McCance, R. A. and E. M. Widdowson. *Proc. Roy. Soc. B.* 138: 115, 1951.
18. Hardy, J. D., P. K. Sen and D. L. Drabkin. *Surg., Gynec. & Obst.* 93: 103, 1951.
19. Mees, C. E. K. and J. R. Baker. *The Path of Science.* New York: Wiley, 1946.
20. Dublin, L. I., A. J. Lotka and M. Spiegelman. *Length of Life.* New York: Ronald Press, 1949.
20a. Keys, A. *J. Am. Dietet. Assoc.* 24: 281, 1948.
21. Brožek, J. and A. Keys. *Science* 112: 788, 1950.
22. Brožek, J. and A. Keys. *J. Gerontol.* 6: Suppl. No. 3, 67, 1951.
23. Sarkisian, S. S. *U. S. Naval Med. Bull.* 46: 1207, 1946.
24. Unpublished data, Lab. of Physiol. Hygiene, Univ. of Minnesota, Minneapolis.
25. Brožek, J. and A. Keys. *Brit. J. Nutr.* 5: 194, 1951.
26. Brožek, J. and A. Keys. *J. Applied Physiol.* 2: 240, 1949.
27. Keys, A. *Chicago Heart Assoc. Bull.* 26: 3, 1948.
28. Keys, A., J. Brožek, A. Henschel, O. Mickelsen and H. L. Taylor. *The Biology of Human Starvation.* Minneapolis: Univ. of Minn. Press, 1950.
29. Osserman, E. F., G. C. Pitts, W. C. Welham, and A. R. Behnke. *J. Applied Physiol.* 2: 633, 1950.
30. Osserman, E. F. Personal communication. 1952.
31. Brožek, J. *J. Gerontol.* 7: 410, 1952.
32. Taylor, H. L. and A. Keys. *Science* 112: 215, 1950.
33. Boothby, W. M., J. Berkson and H. L. Dunn. *Am. J. Physiol.* 116: 468, 1936.
34. Henschel, A. *Federation Proc.* 10: 62, 1951.
35. McCay, C. M. *Vitamins and Hormones* 7: 147, 1947.
36. von Bezold, A. Z. *Wissensch. Zool.* 8: 487, 1857.
37. Brody, S. *Bioenergetics and Growth.* New York: Reinhold, 1945.
38. *Recommended Dietary Allowances (Revised 1948),* Food and Nutrition Board, National Research Council. Washington, D. C.
39. *Calorie Requirements.* Committee on Caloric Requirements, Food and Agriculture Organization of the United Nations. Washington D. C., 1950.

FUNDAMENTAL CONSIDERATIONS, OLD AND NEW, IN NUTRITION OF THE ELDERLY

FREDERIC D. ZEMAN

From the Medical Department, Home for Aged and Infirm Hebrews, New York City

SOME years ago, in the course of studies of the medical history of old age, I was somewhat startled to encounter in a 15th century Italian work a reference to the use of human milk in feeding the aged (1). The author, Gabriele Zerbi, a distinguished physician of Padua and Rome, advised this measure in such a casual and detailed way that it clearly was a well established method of treatment. Subsequent search of the literature disclosed its legendary origin in the story of Cimon, an early Christian, condemned to death by starvation, who was nourished in prison by the breasts of his devoted daughter, Pero. Many artists, including Peter Paul Rubens, have portrayed the bearded father in the act of taking nourishment, under the title of "Filial Piety."

Restrictions of time, as well as a sense of obligation to my principal theme, prevent going into the various details of efforts to track down this fascinating practice, which probably has its roots in remote antiquity and is based on the obvious analogy between the weakness of the baby and the feebleness of old age. I may be permitted however, to remind you that Edmund Burke in his famous speech, *On Conciliation with the Colonies* (March 22, 1775), used it as an effective figure in pointing out that Great Britain had been fed from the New World: "The scarcity which you have felt would have been a desolating famine if this child of your old age, with a true filial piety, with a Roman charity, had not put the full breast of its youthful exuberance to the mouth of its exhausted parent" (2).

Two modern books also mention this method of alimentation. In O'Connor's *The Astors* the author describes the 84-year-old John Jacob Astor as he lay moribund in 1945: "... despite the golden flow into his coffers, his palsied hands must seek the breast of a wet nurse for his nourishment" (3). The closing chapter of Steinbeck's *The Grapes of Wrath* portrays the starving old man nourished by a girl who has just lost her baby (4).

This little known use of human milk illustrates in both its practical and symbolic implications the thinking of the medieval physician in attempting to solve the problems of his aged patients. In Avicenna's *Canon*, (ca. 1000 A.D.), which has been described by Garrison as a "huge unwieldy storehouse of learning," we find an entire chapter devoted to "General Remarks on the Regimen of Old Age," in which a detailed discussion of the food for old people is to be found (5). Judging by the frequency with which they have been quoted, these rules had a wide influence on later writers for hundreds of years, and because of their common sense have validity today, however antiquated the underlying reasons may have become. The following paragraphs are typical:

> Food should be given in small amounts at a time. There may be two or three meals a day, divided up according to the digestive power, and according to the general condition whether robust or weakly. In the latter case at the second or third hour they may partake of well baked bread and honey. At the seventh hour after the bath they may partake of some one or another of the foods we shall name later which are laxative in action. At bedtime some laudable nutriment may be allowed.
>
> When they are robust, old persons may have a rather more liberal supper as long as they avoid any gross aliment which is likely to give rise to atrabilious or serous humor, and avoid all hot, sharp or desiccative foods, such as dishes made with vinegar, salt or hot aromatics, seasoning or pickles. These may, however, be allowed as medicaments.

These examples of ancient medical practice remind us that students of history have no difficulty in furnishing evidence to show that most modern scientific discoveries were anticipated by someone, somewhere in the distant past. Cornaro (1464–1566) vividly demonstrated the value of undernutrition on himself more than three and a half centuries before Dr. McCay (who

is not to blame for the delay) could get around to devising standard experimental conditions by which the life span of white rats is regularly doubled through the use of diets planned to retard growth. Cornaro's close observation of himself is in the true Renaissance spirit, as contrasted with the usage of fifteen centuries, whereby all study was concentrated on the "Great Books," namely, the works of Galen and his commentators. McCay's application of the experimental method to the problems of longevity and of geriatric nutrition exemplifies the spirit and the technique of modern investigation, using the combined efforts of a team of specialists. I need not remind this audience of the achievements of scientific nutrition, nor do I intend to review the successful application to old age of the studies of the role of various food stuffs, of vitamins, of minerals and of their complicated interaction with the body as a whole and particularly with the various endocrine glands.

To physiologists and clinicians concerned with the nature of old age in human beings, the rat colony must always remain an unattainable ideal, in which all variables can be controlled or modified at will,—a condition obviously and lamentably impossible with our patients. Dr. Keys, in discussing the influence of diet on degenerative diseases, has emphasized the difficulties of experiments on man and points out that "to an experimentalist it is distressing to have to admit that major reliance must be placed on methods which are non-experimental, or at least provide no experimental control" (6). My remarks today are concerned with certain basic difficulties, hitherto not sufficiently appreciated, which are inherent in studying elderly individuals, and will be based on clinical and pathological experience at the Home for Aged and Infirm Hebrews, as well as on special studies carried out by members of my staff.

The physician of today in studying old people attempts to differentiate between changes which may properly be ascribed to the aging process itself, and those which are the result of superimposed diseases. Over the years there has been a steady transfer of conditions from the first to the second category. More than fifty years ago Parkes Weber, in the first edition of the *Allbutt-Rolleston System of Medicine*, pointed out that there had been a time when the tuberculous scars commonly seen at the apices of the lungs were ascribed to "senile pneumonitis". Today we are still engaged in investigations designed not only to determine the nature of structural changes in the old but also to clarify the modifications of physiologic responses in the aging organism. In setting up our plans of procedure we must study each individual with care to avoid including any with diseases that might alter the basic responses. We must exercise caution in utilizing old people classified only according to chronological age. We must not be deluded by apparent or outward health, since older people, as we have pointed out elsewhere, show a great contrast between the diagnosable pathological lesions in their bodies and their functional capacity. Often their ability to carry on a lively existence in spite of physical handicaps is most surprising. For this reason we have long followed the example of the cardiologists and added a functional diagnosis to our listing of abnormal conditions found on physical examination (7).

One hardly needs to emphasize that the investigation must be planned in such a way as to prevent any deleterious effect upon the subject, but I have often wondered what changes are produced in certain types of studies by the pain, fear and anxiety caused by repeated venipunctures. Numerous investigators have studied glucose tolerance in the aged, and have reported decreased tolerance. Others, including Horvath, Wisotsky and Corwin (8), and Dr. Robert Newburger (9), of my staff, have been impressed by the variability of repeated tests. Aside from other possible influences on blood sugar responses, I believe that some of the changes observed from one test to another are due to fear and anxiety, shown long ago by Cannon as having potent influences on blood sugar mobilization.

The emotions have often been neglected in evaluating the responses of aged patients. I wish to remind you of the work of Dr. Harold Wolff and his group at the New York Hospital on the effects of emotional stresses of all kinds on the gastrointestinal tract (10), on the cardiovascular system (11) and on the mechanisms of headache (12). Cobb's small volume, *Emotions and Clinical Medicine*, is a splendid review of this whole field (13). While these studies are concerned with all age groups, they are particularly relevant to the complex problems of our elderly patients, in whom we commonly see combined longstanding psychoneuroses, psychosomatic disorders, the changes of senescence, and organic disease of many kinds. These complicated problems not only present clinical difficulties but have pro-

found, often intangible, effects on our physiologic inquiries.

In a recent study of peptic ulcer in the aged Dr. Roger Steinhardt of my staff was surprised to find a low incidence of this disease in the residents at the Home for Aged and Infirm Hebrews, as compared with its frequency in a general hospital outpatient department (14). He believes that this disparity is to be explained on the basis of the security and freedom from fear enjoyed in a protective institutional setting as compared with the emotional stresses and the frustrations of life in the community, with limited economic resources. He is at present engaged in putting these opinions on a sound statistical basis.

To illustrate these points further I wish to call to your attention our changing viewpoints on colonic function in older individuals. That nearly all old people suffer from constipation is one of the sterotypes of clinical medicine and an important part of what I like to call the popular mythology of old age. Using a simple questionnaire supplemented by special physical and x-ray examinations, Dr. Steinhardt studied a group of the ambulatory residents and found that more than 60 per cent were not constipated, and that of the group that complained of this condition, more than 70 per cent had been constipated most of their lives (15). This kind of sample yields more instructive information than that afforded by individuals consulting gastroenterologists, and shows that constipation, though a common complaint, has little to do with aging. When it develops in old age it is an ominous sign and demands speedy investigation. The patients who have had constipation all their lives are suffering in all probability from the late results of childhood toilet-training. In this connection the psychiatrists have long known of the relationship between depressed states and sluggish bowel action, not infrequently being able to predict the onset of a depressed period by the earlier appearance of retarded colonic activity. We must also, in clarifying our thinking about this common problem, recall the convincing reasons that are commonly advanced to explain constipation in the aged: weakness of the smooth muscle of the bowel, atrophy of the abdominal muscles and changes in the nerves and nerve endings serving the reflex arcs. These facile explanations should always be suspect by reason of their very facility, and above all because of their common acceptance and oft-repeated citation in the literature of the subject.

In order to complete the picture I have drawn of the manifold clinical difficulties that beset the investigator studying the nature of the aging process in relation to nutrition and other problems, I wish to emphasize that with old people the last pieces of evidence are furnished by the pathologist. A physician may go through a long life of activity convinced of his diagnostic acumen, provided he never sees post-mortem examinations of his patients. In our institution over 80 per cent of the fatal cases are subjected to necropsy, with results that are often embarrassing but always educational. I recall an individual who had been under our care for over twelve years, regarded because of his quiet unresponsive demeanor as a man of retarded mentality. One night about five years ago, he was missing from his bed and was found to have committed suicide by hanging. Post-mortem examination disclosed a large carcinoma of the stomach. The patient had never given any sign of gastric discomfort. It seems likely that he felt very badly indeed, but because of his mental disease, was unable either to localize his distress, or to tell others about it.

We have also been impressed with the frequency in our residents of asymptomatic myocardial infarction leading to myocardial rupture and sudden death. Here too the presence of mental disease interferes with the patient's understanding and verbalization of his own disturbed functions.

In presenting to you these random historical and clinical notes, bearing not too closely on the topic of the symposium, I have been actuated by the desire to call attention to some of the difficulties inherent in clinical studies of old people, and to stimulate interest in factors that may be overlooked in our tendency toward over-simplification. As I have studied recent literature in this connection, I have been particularly impressed by a brief note published about a year ago by Drs. Brozek and Keyes on *Limitations of the Normal Body Weight as a Criterion of Normality*. They emphasized that "the concept of 'normal' body weight, defined actuarially, tends to obscure the profound changes taking place in the process of aging." They believe that this process involves not only the accumulation of body fat but also the replacement of muscle and other tissues by fat, and point out the im-

plications of their work (16). These investigators are leading the way to new viewpoints based on sound techniques which will surely eliminate many of our current misconceptions and uncertainties. Studies by others on the distribution and amount of body fluids in old age are also of fundamental significance.

To the clinician who has long been puzzled by the manifestations of old age and to the best of his ability has tried to define the details of disease pictures in older people, the growing interest in these problems in laboratories all over the country makes certain our steady progress toward increased understanding. Of highest necessity is the cooperation of biochemist, physiologist and clinician in the combined application of their special skills to these difficult but ever-pressing problems. Only in this way can we hope to shake off the accumulated errors of the past and create a sounder foundation upon which our successors will, we trust, build far more wisely than ever before.

REFERENCES

1. ZEMAN, F. D. *J. Mt. Sinai Hosp.* 10: 710, 1944.
2. BURKE, E. *Speeches and Letters on American Affairs. Everyman's Library, No. 340.* London: J. M. Dent and Sons, Ltd., 1908.
3. O'CONNOR, H. *The Astors.* New York: Knopf, 1941.
4. STEINBECK, J. *The Grapes of Wrath.* New York: Viking Press, 1939.
5. GRUNER, O. C. *A Treatise of the Canon of Medicine of Avicenna* (incorporating a translation of the first book). London: Luzac and Co., 1930.
6. KEYS, A. *J. Am. Dietet. Assoc.* 24: 281, 1948.
7. ZEMAN, F. D. *J. Mt. Sinai Hosp.* 14: 721, 1947.
8. HORVATH, S. M., R. WISOTSKY AND W. CORWIN. *J. Gerontol.* 2: 25, 1947.
9. NEWBURGER, R. Personal communication.
10. WOLF, S., AND H. G. WOLFF. *Human Gastric Function. An Experimental Study of Man and His Stomach.* New York: Oxford Univ. Press, 1947; GRACE, W., S. WOLF, AND H. G. WOLFF. *The Human Colon.* New York: Hoeber, 1951.
11. WOLFF, H. G. *Circulation* 1: 187, 1950.
12. WOLFF, H. G. *Headaches and Other Head Pains.* New York: Oxford Univ. Press, 1948.
13. COBB, S. *Emotions and Clinical Medicine.* New York: Norton, 1950.
14. STEINHARDT, R. Unpublished observations.
15. STEINHARDT, R. Unpublished observations.
16. BROŽEK, J. AND A. KEYS. *Science* 112: 788, 1950.

AMERICAN ASSOCIATION OF IMMUNOLOGISTS

Symposium on Antigenic Variation of Influenza Viruses and Its Importance in Vaccination[1]

Chairman: T. P. MAGILL

A PATTERN OF ANTIGEN VARIATION

M. R. HILLEMAN

From the Army Medical Service Graduate School, Washington, D. C.

VERY early in the history of serological studies on influenza virus it was shown that there was antigenic heterogeneity among the A type strains (1–3). It was found, in addition, that this polyvalency resulted from both qualitative and quantitative variations in a number of ill-defined and irregularly distributed components which formed the antigenic mosaic of the virus.

The early attempts (1–3) to give precise definition and measurement to the varying antigenic composition of these viruses were hampered by the laborious nature of the methods which were then available. Classifications based on the results of serum neutralization and cross-protection tests attempted segregation of the strains into well-defined subtypes but the strains in the various groups tended to merge into each other. Further, there was poor correlation in the conclusions reached in the separate studies by the various workers.

During the early post-war period a special set of circumstances developed which gave impetus to a reexamination of the antigenic patterns of influenza viruses. First, the polyvalent vaccine employed in 1947 failed to protect against the antigenically different A type viruses which became prevalent at that time (4–10). Second, the entrance of the World Health Organization into the influenza field permitted an integrated study of this disease on a global basis for the first time, and made strains available for study from all parts of the world. Third, the adoption of Hirst's hemagglutination-inhibition method (11) for quantitating influenzal antibody provided a simple and rapid procedure for strain analysis. The hemagglutination-inhibition method found widespread use and largely replaced the techniques employed heretofore.

A number of laboratories became interested in the influenza problem and undertook to analyze these agents. Even though most of the comparisons were made by the hemagglutination-inhibition method, there was no uniformity in the tests and this made for difficulty in comparing results. Thus, the use of several kinds of animal sera resulted in a differing degree of strain specificity. Some of the sera were treated to remove non-specific inhibitory substances and others were not. There was a somewhat random selection in the strains compared in the individual studies. Finally, most of the older strains had been through a great many mouse passages which, as shown by Hirst (12), alters the antigenic composition of influenza virus. Although these differences gave rise to variable results in the specific sense, there was nevertheless a wide area of agreement in the broader aspect.

I should like at this time to present the results of some of our studies (13–15) on the influenza viruses, and to let them serve as a basis for a more general discussion of the strain relationships as they appear to be at the present time. All the analyses in our laboratory were made with chicken antisera, since these animals were found to yield sera of sufficiently high titer to permit use at dilutions exceeding the level of non-specific inhibitor. A further advantage of chicken serum was in the relatively uniform antigenic response from bird to bird immunized with the same virus, which was in contrast to the variable antigenic spread which has been reported to result from infection in ferrets (16–17). The hemagglutinating antigens consisted of dried infected allantoid fluid from embryonated eggs (18).

[1] New York City, April 16, 1952.

A typical example of results of analyses of B type influenza viruses is shown in table 1. Antisera to the viruses indicated in the left vertical column were tested with the viral antigens listed in the top horizontal line. For purpose of simplification, the results are given in interpretive form in which the extent of relationship between two viruses is expressed in terms of the number of plus marks. By this method, a 4+ designation for the reaction between heterologous antigen and antiserum indicated that the heterologous virus was inhibited to the same extent as the homologous and 0 indicated complete lack of cross-reaction. Intermediate reactions were indicated by the 3+, 2+, 1+, or ± designations. It is evident in the slide that the B type viruses examined fell into two general groups based on antigenic pattern, and resembled either the Lee strain or the Warner virus. There were minor antigenic variations within the Warner group.

TABLE 1. INTERPRETIVE SUMMARY OF HEMAGGLUTINATION-INHIBITION TESTS SHOWING RELATIONSHIPS OF INFLUENZA B STRAINS

ROOSTER ANTISERUM	ANTIGEN						
	Lee, 1940	Hawaii 1944–45	Warner, 1948	Pooh, 1950	MB, 1950	Meguro, 1950	L1-B-50, 1950
Lee, 1940	4+	1+	±	1+	±	1+	±
Hawaii, 1944–45	2+	4+	1+	2+	2+	3+	3+
Warner, 1948	0	4+	4+	4+	3+	4+	3+
Pooh, 1950	0	4+	4+	4+	3+	4+	3+
MB, 1950	1+	3+	±	1+	4+	2+	3+
Meguro, 1950	0	2+	3+	4+	2+	4+	3+
L1-B-50, 1950	2+	4+	2+	3+	3+	4+	4+

These findings with chicken antisera are in close agreement with the observations of other workers (17, 19–25). There is general concurrence in the antigenic dissimilarity of the 1940 Lee virus from essentially all the strains isolated since that time. In addition, the available evidence suggested that the post-1940 B viruses should be considered a single subgroup, even though numerous minor antigenic differences were evident. In a recent examination of 29 B type viruses (17), Brans in Holland found two strains which were antigenically different from those of the two groups already mentioned. The one virus, TM, occurred in 1940 and the other, Paddington, was isolated in 1943. It should be pointed out in this connection that the TM and Paddington strains, as well as the Lee virus, had all undergone passages in ferrets and mice while essentially all the remaining B viruses were maintained by egg passage alone; thus, the passage history may have been responsible in part at least for the antigenic difference exhibited by these viruses.

The results of similar analyses with chicken serum of A type viruses which were recovered during the period from 1933 to 1951 is shown in table 2. Unlike the B type agents, these viruses comprised a graded series of antigenic forms in which precise subgrouping was impossible. While total antigenic dissimilarity was evident on the far ends of the table, the antigenic patterns of the individual strains blended into each other. Workers who have used sera which were less strain specific than that of the chicken have stated (26–29) that they were able to divide the A type viruses into four general subgroups. The first was the swine influenza viruses, which were not examined in our tests. Second, was the WS virus which occurred in 1933 and showed only partial relationship to the PR8 strain. Third was the PR8-like viruses which were prevalent during the period 1934–1945. These PR8-like agents were very heterogeneous antigenically, but they were sufficiently different from the other A strains to provide a rationale for including them in a single subgroup. Most of the viruses in this group were passed many times in mice, and this may have been responsible for at least some of the wide antigenic diversity observed. The fourth category was the so-called A prime group which has been prevalent since 1946. Within this group of viruses, which were passed almost exclusively in eggs, there was some heterogeneity resulting from an apparent antigenic shift which took place since 1947.

The extent and course of the antigenic change which occurred in the A prime group and was revealed in tests with chicken antiserum is illustrated in table 3. The examination of a number of strains recovered between 1947 and 1950 revealed two significant facts. First, there was marked antigenic homogeneity among the strains which occurred in any particular year, and second, a marked antigenic change over the four-year

period was shown. This change appeared to be a progressive one suggesting that a continuous selection of antigenic variants rather than a discontinuous mutation had taken place in nature. The observations of Taylor (30) and of Archetti and Horsfall (31) have a bearing in this connection. These workers found that passage of an influenza virus in the presence of antiserum to a heterologous but related strain resulted in the selection of antigenic variants which differed markedly from the initial parent strain; the variant viruses bred true on further passage in the absence of immune serum. It is tempting to speculate that the antigenic change which took place naturally in the A prime group since 1947 may have been the result of a similar selection of variants during passage in the partially immune human host, and that this may be at least one of the mechanisms for the appearance of new antigenic forms of A type influenza virus. It is of further importance to point out that while the more recent A prime strains shown in table 3 failed to react with antiserum against those isolated earlier, the newer agents did, nevertheless, retain the ability to stimulate antibody to the older forms. This suggested a successive modification of the surface of the virus, apparently without demonstrable loss of its previous antigenic structure during the time period of this observation. The importance of altered antigenic surface in terms of the survival of the parasite in a partially immune host population has been recognized by Burnet, who suggested that (28) "influenza virus survives by constantly building serological novelty into its past antigenic structure."

While it has been shown (see also 22, 28, 32, 33) that influenza A virus seems to be changing antigenically from time to time, it is of equal importance to recognize that the strains of virus occurring in any particular outbreak tend to be relatively homogeneous antigenically, provided they have not been altered through animal passage (see also 22, 34; cf. 32). The apparent homogeneity observed in single outbreaks might well be extended to include the bulk of the strains occurring in any given year (see also 22, 34). This was evident in the results of analyses by the hemagglutination-inhibition method performed with strains isolated from the epidemic of 1950–51. The findings with a few representative strains are shown in table 4. It is readily apparent that the 1950–51 viruses which were recovered during an approximate 12-month period from widely separated areas of the world were nonetheless relatively homogeneous antigenically and closely resembled the FW-1-50 strain. There was one notable exception in that the Benson virus, isolated at Pittsburgh in 1951, was essentially identical with the 1947 FM1 agent and differed from the remaining 1950–51 strains.

Isaacs and Andrewes (29, 32) have recently stated that they were able to subdivide the 1950-

TABLE 2. INTERPRETIVE SUMMARY OF HEMAGGLUTINATION-INHIBITION TESTS SHOWING RELATIONSHIPS OF INFLUENZA A STRAINS

ROOSTER ANTISERA	ANTIGENS											
	WS, 1933	PR8, 1934	Weiss, 1943	Henry, 1936–37	Hickcox, 1940–41	Hyde, 1935	Hume, 1947	FM1, 1947	Wilfong, 1948	SF-1-49, 1949	FW-1-50, 1950	FM-10-51 1951
WS, 1933	4+	3+	2+	±	1+	0	0	±	0	0	0	0
PR8, 1934	0	4+	2+	4+	±	3+	2+	0	0	0	0	0
Weiss, 1943	0	2+	4+	2+	1+	2+	1+	0	0	0	0	
Henry, 1936–37	±	3+	3+	4+	2+	4+	3+	1+	1+	0	0	
Hickcox, 1940–41	2+	4+	4+	3+	4+	4+	3+	2+	2+	1+	±	
Hyde, 1935	0	1+	2+	4+	2+	4+	3+	1+	1+	0	0	
Hume, 1947	0	1+	3+	3+	3+	4+	4+	1+	3+	0	0	
FM1, 1947	0	±	2+	3+	3+	4+	3+	4+	4+	±	±	1+
Wilfong, 1948	0	1+	2+	2+	2+	2+	2+	3+	4+	2+	2+	
SF-1-49, 1949	0	±	0	1+	1+	1+	1+	4+	3+	4+	3+	
FW-1-50, 1950	0	±	0	2+	2+	2+	0	3+	4+	4+	4+	4+
FM-10-51, 1951	0	±						4+			3+	4+

5[illegible] strains into 2 basic subgroups which they have designated as L and S forms. The principal basis for this division was in the relative avidity of a given strain for its homologous antiserum prepared in ferrets. Thus, it was reported that the S strains were poorly neutralized by the homologous serum even though the same serum contained considerable antibody for a heterologous virus. The L forms, by contrast, demonstrated high avidity for homologous antiserum. It was suggested that the L and S forms might represent alternate phases of one basic subtype of virus, the S phase being the form by which the virus is best able to exist in a population of some basic immunity. While this hypothesis regarding the biological significance of the L and S forms must still be subjected to further observation, it should be recognized that L and S differentiation may provide another useful tool for strain analysis as applied to studies on epidemiology in this disease. It should be added that, to the present time, the difference in avidity of the L and S strains has been observed only in tests with ferret antisera and has not been evident when chicken sera were examined.

TABLE 3. INTERPRETIVE SUMMARY OF HEMAGGLUTINATION-INHIBITION TESTS SHOWING RELATIONSHIPS OF TYPE A AND TYPE A PRIME STRAINS ACCORDING TO YEAR OF ISOLATION

ROOSTER ANTISERA FROM STRAINS ISOLATED IN YEARS	ANTIGENS FROM STRAINS ISOLATED IN YEARS					
	1933	1943	1947	1948	1949	1950
1933	4+	1+	1+	0	0	0
1943	2+	4+	1+	1+	0	0
1947	1+	1+	4+	3+	1+	1+
1948	1+	1+	4+	3+	2+	1+
1949	0	0	4+	4+	4+	4+
1950	0	0	4+	4+	4+	4+

Knowledge of the antigenic patterns of strains of influenza virus is of importance both in epidemiological investigations of the disease and in preventive medicine when prophylactic vaccination is employed. To afford maximal usefulness, rapid methods for analysis should be developed which provide definitive information regarding the antigenic composition of the virus in question and its relationship, not only to previously occurring strains, but to the viruses which are included in the current formula vaccine. In any attempt to measure the antigenic qualities of a given virus, it would appear desirable, *a priori*, to define the antigenic constitution of that virus in terms of the quality and amount of each antigenic component as suggested by Friedewald (35). This theoretical ideal appears impracticable of attainment at the present time. It is fortunate in the more practical sense, however, that such precise measurement may not be necessary to provide the information which might prove useful in the control of this disease by the methods which are presently available. With this philosophy in mind, we proposed a rather simple system (15) which has proved useful for designating the antigenic constitution of influenza viruses.

TABLE 4. INTERPRETIVE SUMMARY OF HEMAGGLUTINATION-INHIBITION TESTS SHOWING RELATIONSHIPS OF 1950–51 INFLUENZA A STRAINS

ROOSTER ANTISERUM	ANTIGEN							
	PR8, 1934	FM1, 1947	FW-1-50, 1950	SWE-3-50, 1950	Lond-1-51, 1951	Eire-1-51, 1951	Bur 99, 1951	Gree-3-51, 1951
PR8, 1934	4+	0	0	0	0	0	0	0
FM1, 1947	0	4+	±	1+	±	±	1+	1+
FW-1-50, 1950	0	4+	4+	5+	3+	3+	4+	4+
SWE-3-50, 1950	0	1+	2+	4+	3+	4+	4+	4+
Lond-1-51, 1951	1+	4+	3+	4+	4+	3+	4+	4+
Eire-1-51, 1951	0	4+	3+	4+	4+	4+	4+	4+
Bur 99, 1951	0	4+	2+	4+	4+	4+	4+	5+
Gree-3-51, 1951	0	3+	3+	3+	3+	2+	4+	4+

TABLE 5. ANTIGENIC FORMULAE OF SOME INFLUENZA A, B AND C VIRUSES

STRAIN	FORMULA (*WS*) (*PR8*) (*FM1*) (*FW1-50*)	STRAIN	FORMULA (*Lee*) (*Warner*)
WS	$Aa_3b_1c_0d_0$	Lee	Ba_3b_0
PR8	$Aa_2b_3c_0d_1$	Warner	Ba_1b_3
FM1	$Aa_1b_0c_3d_3$	IB1	Ba_0b_3
Wilfong	$Aa_0b_0c_3d_3$		
SF-1-49	$Aa_0b_0c_1d_3$		(*1233*)
FW-1-50	$Aa_0b_0c_1d_3$	1233	Ca_3
London-1-51	$Aa_0b_0c_1d_3$	JJ	Ca_3

The system (15) which was developed was founded on the observation that all the A type viruses reacted to a marked extent with the chicken antisera prepared against one or more of four selected strains, i.e., WS, PR8, FM1 and FW-1-50. Similarly, all the B viruses examined reacted with one or both of the antisera against

the Lee and Warner agents. The antigenic formulations were based on the extent of reaction of the virus under study with the antisera to the prototype strains. Representative formulae based on the average results of a number of tests are given in table 5. In the formula for the A viruses, the capital letter A designated the virus type and the small letters 'a,' 'b,' 'c' and 'd' referred to the prototype antisera WS, PR8, FM1 and FW-1-50. The extent of reaction of the virus with each serum was designated 3, 2, 1 or 0 indicating strong, moderate, weak or no inhibition of hemagglutination. Thus, for example, the formula $Aa_2b_3c_0d_1$ is given for the PR8 strain. The essential difference of WS from PR8 and of the dissimilarity of both of these strains from the A prime group is clearly evident. All the 1947 and 1948 A prime viruses examined had the formula $Aa_1b_0c_3d_3$ or $Aa_0b_0c_3d_3$ and the 1949, 1950 and 1951 strains bore the designation $Aa_0b_0c_1d_3$. The B type agents which we examined were similar either to Lee which is called a, or to Warner signified by the letter b. B viruses which occurred since 1943, were designated Ba_1b_3 or Ba_0b_3. The work of Brans (17) would indicate the need for 2 additional prototypes but since these were represented by only a single example in each case and since both the viruses were recovered prior to 1944, it is questionable whether there is a real need to include them in analyses of the strains occurring at the present time. The influenza type C viruses (36), which were not mentioned earlier, were totally homogeneous antigenically and a single prototype was sufficient to cover all the strains tested to date.

The importance of major variation in antigenic pattern of influenza viruses in its relationship to the vaccination problem was clearly established (4–10) in 1947. It was shown that vaccine containing the PR8 (1934) and Weiss (1943) strains failed to give protection against the so-called A prime agents which became prevalent at that time. The effect of variation of less extent upon vaccine effectiveness remains to be clarified. It should be pointed out in this connection, however, that a monovalent vaccine prepared from the 1947 FM1 virus was effective in reducing the incidence of A type influenza during an epidemic in 1950 (37). Laboratory analyses of the antigenic patterns of influenza viruses may assist materially in the selection of strains for inclusion in a vaccine. The choice of such strains should not, however, be based on laboratory studies alone and final proof of usefulness must await the results field trials in man.

REFERENCES

1. Magill, T. P. and T. Francis, Jr. *Brit. J. Exper. Path.* 19: 273, 1938.
2. Francis, T., Jr. and T. P. Magill. *Brit. J. Exper. Path.* 19: 284, 1938.
3. Smith, W. and C. H. Andrewes. *Brit. J. Exper. Path.* 19: 293, 1938.
4. Francis, T., Jr., J. E. Salk and J. J. Quilligan, Jr. *Am. J. Pub. Health* 37: 1013, 1947.
5. Loosli, C. G., J. Schoenberger and G. Barnett. *J. Lab. & Clin. Med.* 33: 789, 1948.
6. Mellanby, H., C. H. Andrewes, J. A. Dudgeon and D. G. MacKay. *Lancet* 1: 978, 1948.
7. Dudgeon, J. A., H. Mellanby, R. E. Glover and C. H. Andrewes. *Brit. J. Exper. Path.* 29: 132, 1948.
8. Sigel, M. M., F. W. Shaffer, M. W. Kirber, A. B. Light and W. Henle. *J. A. M. A.* 136: 437, 1948.
9. Rasmussen, A. F., J. C. Stokes and J. E. Smadel. *Am. J. Hyg.* 47: 142, 1948.
10. Salk, J. E. and P. C. Suriano. *Am. J. Pub. Health* 39: 345, 1949.
11. Hirst, G. K. *J. Exper. Med.* 75: 49, 1942.
12. Hirst, G. K. *J. Exper. Med.* 86: 357, 1947.
13. Hilleman, M. R., R. P. Mason and N. G. Rogers. *Pub. Health Rep.* 65: 771, 1950.
14. Hilleman, M. R., R. P. Mason and E. L. Buescher. *Proc. Soc. Exper. Biol. & Med.* 75: 829, 1950.
15. Hilleman, M. R. *Proc. Soc. Exper. Biol. & Med.* 78: 208, 1951.
16. Hirst, G. K. *J. Exper. Med.* 78: 407, 1943.
17. Brans, L. M. *Onderz. en. Med. Uit. Het. Inst. v. Praev. Genees. No. 9.* Leiden: Stenfert Kroese, 1952.
18. Hilleman, M. R., E. L. Buescher and J. E. Smadel. *Pub. Health. Rep.* 66: 1195, 1951.
19. Burnet, F. M. and J. D. Stone. *Australian J. Exper. Biol. & M. Sc.* 24: 207, 1946.
20. Dudgeon, J. A., C. H. Stuart-Harris, C. H. Andrewes, R. E. Glover and W. H. Bradley. *Lancet* 2: 627, 1946.
21. Francis, T., Jr., J. E. Salk and W. M. Brace. *J. A. M. A.* 131: 275, 1946.
22. Hirst, G. K. *J. Exper. Med.* 86: 367, 1947.
23. Lazarus, A. S. *Proc. Soc. Exper. Biol. & Med.* 71: 317, 1949.
24. Tamm, I., E. D. Kilbourne and F. L. Horsfall, Jr. *Proc. Soc. Exper. Biol. & Med.* 75: 89, 1950.
25. Bozzo, A. *WHO Bull.* 5: 149, 1952.

26. ANDREWES, C. H. *Proc. Roy. Soc. Med.* 42: 519, 1949.
27. VAN DER VEEN, J. AND J. MULDER. *Onderz. en. Med. Uit. Het. Inst. v. Praev. Genees. No. 6.* Leiden: Stenfert Kroese, 1950.
28. BURNET, F. M. *Bull. Johns Hopkins Hosp.* 88: 157, 1951.
29. ISAACS, A. *Proc. Roy. Soc. Med.* 44: 801, 1951.
30. TAYLOR, R. M. *Am. J. Pub. Health.* 39: 171, 1949.
31. ARCHETTI, I. AND F. L. HORSFALL, JR. *J. Exper. Med.* 92: 441, 1950.
32. ISAACS, A. AND C. H. ANDREWES. *Brit. Med. J.* 2: 921, 1951.
33. ANDREWES, C. H. *Proc. Roy. Soc. Med.* 44: 803, 1951.
34. CHU, C. M., C. H. ANDREWES AND A. W. GLEDHILL. *WHO Bull.* 3: 187, 1950.
35. FRIEDEWALD, W. F. *J. Exper. Med.* 79: 633, 1944.
36. FRANCIS, T., JR., J. J. QUILLIGAN, JR. AND E. MINUSE. *Science* 112: 495, 1950.
37. MEIKELJOHN, G., C. H. KEMPE, W. G. THALMAN AND E. H. LENNETTE. *Am. J. Hyg.* 55: 12, 1952.

BIOLOGICAL SELECTION OF ANTIGENIC VARIANTS

FRANK L. HORSFALL, JR.

From the Hospital of The Rockefeller Institute for Medical Research, New York City

THAT antigenic variation of influenza viruses is not in all cases an artifact caused by manipulation of these agents in the laboratory has been established decisively. Although variation in antigenic constitution may result from passage in experimental animals which are unnatural hosts for influenza viruses, it has become increasingly clear that similar variation occurs during successive epidemics of the natural disease in human beings. Within three years of the discovery of influenza A virus, there were indications of differences in the antigenic composition of strains of this agent (1). One year after the discovery of influenza B virus the antigenic homogeneity of strains of this agent was questioned (2, 3). These initial observations have been abundantly confirmed and widely extended by numerous workers (4–13).

The degree to which some laboratory strains of these agents differ, and the finding that among most strains detectable differences can be discerned, led some investigators to question the significance of antigenic variation as bearing on the disease in man. Because vaccines prepared with old laboratory strains of influenza A virus continued to give indications of causing increased resistance to influenza A during epidemics occurring from 1941 (14) through 1945 (15, 16), it appeared that antigenic variation, although a phenomenon of considerable theoretical interest, might have but little meaning relative to the natural disease.

Such a view of this aspect of viral variation became untenable during influenza A epidemics in 1947. As a result of the emergence of strains markedly dissimilar from pre-1947 strains (17–19), an event which was heralded by Australian experience (20) in the preceding year, no benefit derived from the use of the old vaccine (17, 18). That antigenic variation of influenza A virus occurred in natural influenza in man could no longer be doubted, and the implications of this phenomenon relative to both natural and artificially induced immunity became strikingly clear. The extent of the variation in antigenic composition which occurred in 1947 was so marked that the new strains were dubbed A-prime (21), an unhappy designation which has led to considerable confusion. Strains recovered since 1947 are not antigenically unrelated to those recovered in earlier years, but in complete cross-serological experiments, the immunological relationship between certain pre- and post-1947 strains appears to be as small as 1 per cent (22).

Antigenic variation of influenza viruses did not cease to occur in 1947, even though the differences observed in that year were more striking than any encountered since. With both influenza A virus and influenza B virus there are clear indications that additional variation in antigenic constitution occurred during the interval between 1947 and 1951 (11, 13, 22). That such differences as have emerged have not yet become sufficiently marked to diminish the effectiveness of vaccines prepared with 1947 strains (23) can hardly be considered as a good reason for thinking that further variation more striking in degree may not occur in the future.

To account for the origin and the biological selection of variants of influenza viruses, it is necessary to make only two assumptions, both of which are supported indirectly by evidence which has been obtained with other infectious agents. These assumptions are: *1*) Viral variants arise spontaneously and continuously; *2*) viral variants are selected for or against by environmental factors in the host which are operative during the multiplication process (24). If some new character provides a distinct preferential advantage for a variant during multiplication in a given host, it would be anticipated that eventually the variant should come to predominate and might actually eliminate the wild type or parent virus. That this expectation is consistent with experience has been shown repeatedly in extensive studies on variation with bacterial viruses (25).

A preferential advantage sufficient to result in selection for an antigenic variant of influenza virus could in theory operate in either of two ways: *1*) Any new character which provided an advantage for the variant in a given host would

favor selection of the variant. A difference in antigenic constitution could be linked to such a new character merely by chance. In a non-immune host, antigenic variants of this kind would tend to be selected because of a non-immunological advantage. *2*) A change in antigenic constitution would of itself provide an advantage for the variant in a partially immune host. A variant which was not neutralized by available antibodies would be selected for. In a partially immune host such antigenic variants would tend to be selected because of an immunological advantage.

The selection of an antigenic variant in an unnatural host in the absence of antibody can be considered as an example of selection on the basis of a new character providing a non-immunological advantage. Excellent evidence that this occurred was obtained in experiments (26) on the mechanism of adaptation of influenza A virus to the mouse. When strains which initially were immunologically indistinguishable were carried through serial lung passages in normal mice, variants appeared which differed in antigenic composition from the parent strains. In addition, the variants differed more widely from each other than from their respective parent strains, indicating a tendency toward the development of divergent antigenic patterns. The extent of the antigenic differences observed was sufficiently marked to make it appear that the mouse passage strains possessed a cross-immunological relationship of only about 18 per cent. The results of these experiments were confirmed readily with other recently recovered strains in this and another laboratory (27).

Because normal mice were employed in the experiments summarized above, it is clear that selection of the antigenic variants was dependent upon factors other than their immunological characters. The mechanism for the selection of the variants, although not yet definitely established, can be understood on the basis that some other new character appeared which afforded a non-immunological advantage. As compared with the parent strain, more efficient adsorption by susceptible cells, a more rapid rate of multiplication in such cells, a higher proportion of infective viral particles, or simply greater stability in the environment provided by the unnatural host, would give the variant a preferential advantage. On serial passage the effects of the advantageous character accumulate and, after a sufficient number of passages, demonstrable differences in the strain become apparent. That an antigenic difference may emerge during this process is incidental, and does not contribute to the selection of the variant virus. However, it appears probable that the new antigenic pattern of the variant is linked to the advantageous character which occasioned selection, for the antigenic composition of the variant tends not to alter further so long as the strain is maintained in the new host. Thus, there are indications that a strain may acquire a relatively fixed antigenic pattern if the host environment is kept constant.

The selection of antigenic variants in the presence of neutralizing antibody can be considered as an example of selection on the basis of a new character providing an immunological advantage. Unequivocal evidence that this occurs with influenza A virus was obtained recently in this laboratory (28). When strains of the agent were passed in series *in ovo* in the presence of neutralizing antibody against another strain, antigenic variants emerged. Such variants possessed antigenic patterns which differed from those of the parent strains and differed even more strikingly from those of the strains used to produce the immune sera. That selection of antigenic variants was occurring under the conditions of these experiments was clearly demonstrable after 3 passages *in ovo* in the presence of heterologous antibody. Variants with new antigenic patterns which persisted through at least 24 passages (29) in the absence of a selective environment, i.e., heterologous antibody, were obtained after only 6 passages.

With strains which were immunologically related, one to the other, to an extent of 15–40 per cent, antiserum against one caused selection of antigenic variants from the other (28). However, if the immunological relationship was of the order of only 2 per cent or less, no selection of variants was demonstrable. As was to be expected, selection of antigenic variants by this procedure depended on the introduction of a large number (i.e., 10^7 or more) of viral particles and a quantity of heterologous antibody sufficient to neutralize a very high proportion of the viral population. This indicates that only a small number of variant particles were present in the original parent population. When neither the strain nor the immune serum was changed, variants with indistinguishable new antigenic patterns were obtained in successive experiments, indicating predictability in selection under constant conditions. With suitable strains and immune sera, selection

of this sort can be accomplished repeatedly and, once antigenic variants have been separated from the bulk of the viral population, their new character tends to persist in the absence of a selective environment. That it is the new antigenic pattern of the variant which provides the preferential advantage and that selection operates on the basis of an immunological advantage seems obvious in this case.

The results of experiments of this sort in laboratory animals makes it possible to propose a concept regarding a mechanism for the selection of antigenic variants during natural transmission of influenza viruses in man. Variants possessing new antigenic patterns which correspond least with the neutralizing antibodies available in partially immune hosts possess an immunological advantage and thus become selected for. This theory can be termed selection on the basis of least fit. Two well established facts provide support for the hypothesis: *1*) Antigenic variants of influenza viruses arise in nature; *2*) normal persons possess neutralizing antibodies against strains with which they have had prior contact.

That there are marked quantitative differences in the antibody levels possessed by normal adults has been demonstrated abundantly in both neutralization (14) and hemagglutination-inhibition (30) tests against certain strains of influenza A virus with serum from large numbers of persons. The extent of the differences in antibody level is remarkably large, greater in fact than has been observed with antibodies in human serum against other viral agents. As determined with the PR8 strain, approximately 80 per cent of normal adults show neutralizing antibodies in their serum, and the titers range from as low as 1:2 to as high as 1:1000 or more (14). Against the same strain, hemagglutination-inhibition is obtained with the sera of over 90 per cent of normal persons in titers which range from 1:32 to 1:1024 or more (30).

The finding that there are qualitative differences (31) in the antibody patterns of normal persons against various strains of influenza A virus is of considerably greater importance to the present thesis. Such differences have been demonstrated recently in this laboratory (32). As an example, some persons who possess high titers of neutralizing antibodies against pre-1947 strains have little or no antibodies against post-1947 strains. In addition strains of influenza A virus recovered from such persons show only a very small immunological relationship to pre-1947 strains. These strains have an antigenic pattern which is in large measure a reciprocal reflection of the antibody pattern of the persons from whom the agents were obtained. When recently recovered strains, passed only in chick embryos, were employed in neutralization tests with acute phase serum from the patients who yielded the agents, antibodies were not demonstrable despite the fact that neutralizing antibodies against antigenically different strains were present, sometimes in high titer.

On the basis of extensive studies of antibodies against influenza viruses present in the serum of normal persons, it appears that most adults may be considered as partially immune hosts. However, their antibody patterns are not uniform and there is reason to think that in most instances they reflect only a restricted segment of the total antigenic potential of influenza viruses. If a large population of viral particles containing a small number of antigenic variants is available to initiate infection in hosts which, although immune to the parent agent, are not immune to the variants, it would be expected that selection of the variants should occur. This may be predicted from the results obtained on partial neutralization of influenza virus *in ovo* with antibody against different strains (28). It appears that infection of adult human beings with this agent may be considered as closely analogous to the *in ovo* antibody model described above, and that the selection of antigenic variants in man is dependent upon a similar immunological advantage.

If this hypothesis relative to a mechanism for the selection of antigenic variants of influenza viruses in human beings has validity, the following predictions may be made: *1*) Minor antigenic differences between strains recovered during a single epidemic year will occur. There are already some indications that such differences are discernible when strains recovered in widely separated geographical areas are compared (22). *2*) More definite antigenic differences will occur among strains recovered in successive years. The results of recent studies conducted in this laboratory (22) provide some support for this statement. *3*) With recurring epidemics in the same population, the longer the time interval between the recovery of strains, the more marked will be the dissimilarity in their antigenic constitution. As yet, an insufficient number of strains has been studied under satisfactory conditions during too short a period of time to provide an adequate test of this possibility.

REFERENCES

1. MAGILL, T. P. AND T. FRANCIS, JR. *Proc. Soc. Exper. Biol. & Med.* 35: 463, 1936.
2. EATON, M. D. AND M. D. BECK. *Proc. Soc. Exper. Biol. & Med.* 48: 177, 1941.
3. GORDON, I. *J. Immunol.* 44: 231, 1942.
4. MAGILL, T. P. AND T. FRANCIS, JR. *Brit. J. Exper. Path.* 19: 273, 1938.
5. FRANCIS, T., JR. AND T. P. MAGILL. *Brit. J. Exper. Path.* 19: 284, 1938.
6. SMITH, W. AND C. H. ANDREWES. *Brit. J. Exper. Path.* 19: 293, 1938.
7. HIRST, G. K. *J. Exper. Med.* 78: 407, 1943.
8. FRIEDEWALD, W. *J. Exper. Med.* 79: 633, 1944.
9. MAGILL, T. P. AND J. Y. SUGG. *J. Exper. Med.* 80: 1, 1944.
10. HIRST, G. K. *J. Exper. Med.* 86: 367, 1947.
11. TAMM, I., E. D. KILBOURNE AND F. L. HORSFALL, JR. *Proc. Soc. Exper. Biol. & Med.* 75: 89, 1950.
12. ISAACS, A. AND C. H. ANDREWES. *Brit. Med. J.* 1: 921, 1951.
13. HILLEMAN, M. R. *Proc. Soc. Exper. Biol. & Med.* 75: 829, 1950.
14. HORSFALL, F. L., JR., E. H. LENNETTE AND E. R. RICKARD. *Pub. Health Reps.* 56: 1863, 1941.
15. Commission on Influenza, Army Epidemiological Board. *J. A. M. A.* 124: 982, 1944.
16. Commission on Influenza, Army Epidemiological Board. (A series of seven papers). *Am. J. Hyg.* 42: 1, 12, 21, 28, 45, 57 and 94, 1945.
17. FRANCIS, T., JR., J. E. SALK AND J. J. QUILLIGAN, JR. *Am. J. Pub. Health* 37: 1013, 1947.
18. SIGEL, M. M., F. W. SHAFFER, M. W. KIRBER, A. B. LIGHT AND W. HENLE. *J. A. M. A.* 136: 437, 1948.
19. RAUSMUSSEN, A. F., JR., J. C. STOKES AND J. E. SMADEL. *Am. J. Hyg.* 47: 142, 1948.
20. ANDERSON, S. G. *Australian J. Exper. Biol. & M. Sc.* 25: 243, 1947.
21. SALK, J. E. AND P. C. SURIANO. *Am. J. Pub. Health* 39: 345, 1949.
22. HILLEMAN, M. R. AND F. L. HORSFALL, JR. *J. Immunol.* In press.
23. MEIKLEJOHN, G., C. H. KEMPE, W. G. THALMAN AND E. H. LENNETTE. *Am. J. Hyg.* 55: 12, 1952.
24. HORSFALL, F. L., JR. *Proc. Second Internat. Poliomyelitis Congress, Copenhagen.* In press.
25. HERSHEY, A. D. *Cold Spring Harbor Symp. Quant. Biol.* 11: 67, 1946.
26. HIRST, G. K. *J. Exper. Med.* 86: 357, 1947.
27. SUGG, J. Y. *J. Bact.* 58: 399, 1949.
28. ARCHETTI, I. AND F. L. HORSFALL, JR. *J. Exper. Med.* 92: 441, 1950.
29. ARCHETTI, I. *Proc. Soc. Exper. Biol. & Med.* 80: 212, 1952.
30. HIRST, G. K., E. R. RICKARD, L. WHITMAN AND F. L. HORSFALL, JR. *J. Exper. Med.* 75: 495, 1942.
31. HORSFALL, F. L., JR. AND E. R. RICKARD. *J. Exper. Med.* 74: 433, 1941.
32. HARDY, P. H., JR. Unpublished observations.

SIGNIFICANCE OF ANTIGENIC VARIATION OF INFLUENZA VIRUSES IN RELATION TO VACCINATION IN MAN[1]

THOMAS FRANCIS, JR.

From the Department of Epidemiology and Virus Laboratory, School of Public Health, University of Michigan, Ann Arbor

ALTHOUGH the significance of antigenic variation in influenza viruses to vaccination has been continuously discussed for the past fifteen years or more, the answer is far from complete. It was promptly demonstrated in the early studies that the virus of swine influenza differed sharply in its antigenic characteristics from the WS and PR8 strains of human origin, but recovery of naturally susceptible animals (ferrets, swine) from infection with either swine or human virus resulted in immunity to large doses of the heterologous virus (1, 2), although recovery was not ordinarily accompanied by significant amounts of antibody to the heterologous virus (3). It was also demonstrated that vaccination of mice with one could establish a definite immunity to the other (3) and the interrelation of WS and PR8 strains was shown by neutralization procedures. All three strains were clearly related by complement-fixing antigen.

At this time we had never worked with the WS strain in our laboratory. It was only after repeated, continued effort that we were able to demonstrate significant antigenic differences among the strains of human origin isolated and maintained in our laboratory. The reason is apparent; they were too much alike, and we commonly employed convalescent serum or serum obtained after repeated inoculations. The establishment of detailed quantitative conditions for cross testing was the other requirement. The first large scale examinations of strains furnished much of the pattern which has been subsequently and repeatedly demonstrated (4–6). It is well to remember that although in 1936 we isolated strains of virus directly in mice, in eggs, and in tissue culture, most of the work was done with strains which had been isolated in ferrets and established in mice. For neutralization tests they were transferred in tissue culture and the sera were derived from rabbits 7–10 days after a single intraperitoneal injection. Some series of neutralization tests were also carried out in tissue cultures. The results clearly demonstrated sharp differences between strains of different years and the tendency for those from a given year to be closely similar. It was also pointed out that practically every strain could be shown to have a distinct antigenic character, if the procedure were maintained at a sufficiently delicate level. The variation in strains was shown also in tests of active cross immunity—an important correlation.

The purpose of this recounting is to emphasize that the problem involves the significance of the demonstrable similarities as well as of the differences. I shall take the position that, from the point of view of vaccine, the mere demonstration that strains differ is no longer of importance. Nor can I dismiss some of this earlier evidence and assume that one can assay the immunological significance of differences by highly selective serological procedures alone. Nor, on the basis of the evidence, can one accept the thesis that results obtained with strains established in mice must be discounted, while those maintained in eggs are highly acceptable. The question is that of significance. Practically all the studies of vaccination with which the Commission on Influenza has been concerned have been directed in large part to a continued evaluation of the significance of antigenic variation in determining the efficacy of vaccine. Of course, it was first necessary to demonstrate that vaccination would induce sufficient immunity in the human population to warrant continued effort along these lines. That evidence was obtained by the 1943 studies of the Commission in which the PR8 and Weiss strains constituted the A component of the vaccine (7). Even here there were minor suggestions that the epidemic strains were not completely identical with those of the vaccine, but the vaccine was effective.

[1] Certain of the studies referred to in this paper were conducted under the auspices of the Commission on Influenza, Armed Forces Epidemiological Board, and supported by the Office of The Surgeon General, Department of the Army, Washington, D. C.

In 1945 the results at Michigan (8) and at Yale (9) with the same vaccine gave evidence that the Lee strain had induced an efficient immunity against the Type B strains of that year. Attention was called to the fact that the protective effect was obtained although the antigenic character of the 1945 strains differed considerably from that of the Lee strain. It is apparent that the results left wide open the question of how different can the viruses be and yet yield cross-protection by vaccination. But they also flatly contradicted any assumption that a serological variation of this degree was of much importance. The differences between the strains were limited in that the vaccinating strains induced good antibody rises to the epidemic ones, indicating that a very important similarity was retained. The old antigen had not disappeared from the new strains.

TABLE 1. ANTIBODY RESPONSE OF A-PRIME PATIENTS IN 1947 TO PR8 AND RHODES (1947 A′) STRAINS[1]

	PR8		RHODES (1947)	
	Titer	Mean Rise	Titer	Mean Rise
Acute	76	3.0	27	4.4
Conv.	230		120	

[1] See Ref. 10.

There was little forewarning of the variant which circulated in 1947, for there had been little influenza A in this country since 1943. Had it been a mutation forced by the immunizing effects of 1943, it might well have been circulated earlier, and in fact, it was subsequently shown to be similar to that prevalent in Australia during 1946. Nevertheless, had it not been for studies in vaccinated populations it seems doubtful that the significance of the variation would have been so readily recognized. While the vaccine of 1946–47 had contained the same strain elements, it was definitely less potent (reduced to N.I.H. requirements) than that employed in 1943 and 1945. As you will recall, there was no evidence of a protective effect. The experience gave important information on a question of many years' asking. But what were the differences which were important?

First of all, the PR8 and Weiss strains failed as antigens to elicit significant antibody to the A-prime strains, although patients recovering from the disease commonly exhibited as good or better antibodies to the PR8 strain as to the epidemic strains.

The data indicate that the A-prime strains possessed a sufficient content of the PR8 component to induce a rise in titer, at least, equivalent to that of A-prime, although the PR8 strain was deficient in the essential A-prime antigen. This conclusion is further substantiated by results obtained in our laboratory in 1948 by the vaccination of men with univalent vaccines of PR8 (A), FM1 (A-prime), and Lee (B), strains of virus.

Here it is seen that vaccination of human subjects with FM1 vaccine induced a parallel response to PR8 and FM1 strains but the PR8 strain stimulated no antibody to the FM1 strain. This is completely in agreement with the data obtained by Meiklejohn, Weiss, Shragg and Lennette (11), who observed a 2.5 increase against PR8 and 3.4 against FM1 after FM1 vaccine. In passing, it also indicates that FM1 is not a potent antigen.

TABLE 2

No. men	Vaccine	TITER VS. PR8 Pre	Post	Incr.	TITER VS. FM1 Pre	Post	Incr.
53	PR8	74	469	6.3	74	120	1.6
54	FM1	64	225	3.5	70	256	3.7
54	Lee	63	75	1.2	63	76	1.2

Hilleman (12) records the comparative antigenic characteristics of the PR8 and FM1 strains as follows:

PR8, 1943	$a_2b_3c_0d_1$
FM1, 1947	$a_1b_0c_3d_3$
Hume, 1947	$a_0b_2c_3d_0$

If this be true, then the 'a' and 'd' components are the antigens responsible for cross relationship, since each is lacking in the primary antigenic component of the other. Actually, the results in man would seem more to indicate that the A-prime strains do possess a fair amount of the primary PR8 antigen, as in the Hume strain. Doctor Magill's classification of the PR8 and FM1 strains shows mutual contents of the respective primary antigens (13). The situation is similar to that of the PR8 and swine strains; the swine strain induces cross reacting antibody with difficulty while the PR8 strain does it readily, suggesting that it is the swine virus which is deficient in the PR8 antigen rather than that the swine antigen has disappeared from the PR8.

The studies with monovalent vaccines at Fort

Ord in 1949–50 (14) have clearly demonstrated that despite the diversity of PR8 and the epidemic strains, there was a mild reduction in the occurrence of influenza among those receiving the PR8 vaccine, suggesting that there was not a complete lack of relationship between the PR8 and the 1949 strains. According to Hilleman they are related only by a one plus crossing of the 1950 component. The effect with the FM1 vaccine was much more pronounced so that among training units there were five times as many cases in the control groups as in those receiving the FM1 preparation. This again, it should be emphasized, appeared to ignore the serological evidence that the 1950 California strain has been shown to deviate from the FM1 strain and to retain similarity largely by overlapping of the 1950 component in the two strains; nevertheless, the mutual relationship is still extensive. Although using these classifications for illustration of the argument, the fact remains that the minutiae are far from uniform in the strains of a sub-group, in different lines of the same strain, or in the results from different laboratories with the same line. It is evident, therefore, that efforts which seek to impose too rapid a conclusion as to their significance will only result in influencing administrative policy rather than in clarifying the problem.

A similar though somewhat less complex situation obtains in the case of Type B strains. The tendency of strains from different years to vary is quite definite and numerous suggestions have been made that preparation to meet this variation must be made by adding new Type B strains to the vaccine. On the contrary, the Commission studies have been limited to the use of only the Lee strain either combined in a vaccine with Type A or in monovalent Type B preparations.

As previously mentioned, strains of virus isolated from the widespread epidemic disease of 1945 differed sharply from the Lee strain of the vaccine. Nevertheless, antibody stimulated in man by the vaccine reacted with the epidemic strains and a staunch protection was afforded Army personnel so that their incidence of disease was one-tenth to one-twentieth that in unvaccinated forces, such as the Navy, under similar conditions of exposure. From 1945 until this year there had been little but scattered flurries of influenza B in this country, most commonly on the West Coast. Certain strains of the virus which had been isolated were found to be different from the Lee strain, and it has been suggested that these strains maintained in eggs are of sufficiently different antigenic makeup as to be needed in the vaccine. It is in a situation such as this that the continued evaluation of vaccine in man has provided the opportunity to gain first hand information as to the significance of strain variation by retaining a single vaccine strain, Lee. The data are not complete, but this year again the evidence is that vaccine has had a pronounced effect upon influenza B. Some have commented that the strains are closer to 1949 strains than to 1945 or Lee. Nevertheless, serologically, most patients have exhibited convalescent rises in titer which are readily detectable with strains of earlier years. Once more it may be suggested that the Type B virus is a better antigen and gives a more solid immunity; hence, moderate variation would have less effect than with Type A. And it can be concluded that a basic, common, immunogenic component of Type B strains has been retained by Type B strains over the past twelve years; the greater interval between widespread Type B epidemics may be a result of that fact. The reported disproportionately high incidence of influenza B in children may be an epidemiologic reflection of a greater strain fixation. Had another strain been added to the vaccine this important information would have been obscured.

The significance of the variation in strains in relation to vaccine has, then, in my opinion, been best estimated to date by actual studies of vaccination. The value of the other procedures which have been advanced for antigenic classification is not yet apparent, since the correlations with immunity have been largely avoided. Moreover, the procedures employed have to a large extent sought to avoid the cross relations and to focus upon the differences, by using sera from insusceptible animals taken after a limited stimulus which may represent only the most ready expression of antigenic constitution. To illustrate this point: in studying relations between different 1947 strains of virus we have observed that mice given two doses of the Keffer strain of virus intraperitoneally exhibited minimal H.I. antibody to the Rhodes strain and yet were actively immune to more than log 5.0 LD_{50} given intranasally.

On the other hand, if data obtained by antibody absorption and by various neutralization or complement-fixation tests can be correlated with immunity tests, they may yield knowledge of the number of distinctive antigens which will be encountered in influenza viruses of a given

type. However, it is difficult at present to embrace all the observations of different workers in one concept, for it is not yet clear that some of the differences are more than serological embroidery; certainly they are not consistent in all hands.

The thesis that changes in influenza virus proceed with succeeding years as an orderly controlled variation based upon the antibody content earlier developed by the population is, to me, an extremely doubtful one, though the report of Archetti and Horsfall appears to give an experimental parallel (15). According to this, earlier strains fall out of existence; a concept somewhat out of keeping with what is known of other microbial agents. But if the concept is correct, within the limits of the Type A or Type B groups, does it not suggest that only a limited number of antigenic components is to be reckoned with? Otherwise, variants of much greater diversity should be anticipated, giving rise to many more types as different as A and B. Given a limited antigenic constitution, the outlook for meeting the problem effectively is quite hopeful. Assume that there are nine antigenic components which may be involved. Something over 360,000 combinations and permutations are possible if only the order of arrangement of these antigens is taken into account. However, if the actual quantity of a given component is the factor which determines the major reactivity, then a combination of nine strains or less providing adequate amounts of the different antigens would meet the situation. There is a great deal to suggest that the dominant antigen may be located at the surface, but even if spatial arrangement is the determining influence, much the same considerations could hold. The extensive studies of the use of practicable adjuvant vaccines now being conducted by Salk and his associates (16, 17) may also meet certain of the difficulties by stimulating much higher, broader and more persistent responses to the different antigens.

One other feature should be of great importance in this discussion—the durability of resistance. If it were true that the antigens of older epidemic strains are progressively less available in the strains of succeeding years, resistance to the former will diminish and a population susceptible to strains of the old varieties will arise. The result should be to force variation back in that direction, thus creating the cycles of recurrence among strains and furnishing a splendid basis for the classical notions of the periodicity of influenza.

The present discussion has been entirely based upon what is commonly called antigenic composition. But the intrinsic factors which give to the virus an epidemic character are not defined. Perhaps they are only combinations of the same antigenic factors we have been measuring, but it seems certain that they have an important place in the selection, from all the variants which can arise, of those capable of rapid dissemination and high pathogenicity (18). The ability of a strain to adapt itself to the substrate offered by the host has, apart from specific immunity, a major significance in the decision as to whether or not a strain will become well established with epidemic characteristics.

The later phase of this talk has been hypothetical and obviously has not discussed all the possibilities which exist. In any case, the evidence gained from studies of vaccination in man seems to me not to justify a defeatist attitude, but to emphasize that the limitation of influenza can be accomplished by vaccination, and on a large scale can be capable of maintaining a more effective immunity of broad character than merely awaiting the influence of recurrent epidemics. It has shown that studies of immunity induced by vaccination to strains of virus in circulation give the most significant data concerning the significance of strain variation to prophylaxis of the disease. The continued study of vaccines and strains will give us thorough knowledge for planning action rather than allowing decisions to be the result of whims and pressures of the moment.

REFERENCES

1. Smith, W., C. H. Andrewes and P. P. Laidlaw. *Brit. J. Exper. Path.* 16: 291, 1935.
2. Shope, R. E. and T. Francis, Jr. *J. Exper. Med.* 64: 791, 1936.
3. Francis, T., Jr. and R. E. Shope. *J. Exper. Med.* 63: 645, 1936.
4. Magill, T. P. and T. Francis, Jr. *Brit. J. Exper. Path.* 19: 273, 1938.
5. Francis, T., Jr. and T. P. Magill. *Brit. J. Exper. Path.* 19: 284, 1938.
6. Smith, W. and C. H. Andrewes. *Brit. J. Exper. Path.* 19: 293, 1938.
7. Commission on Influenza. Studies of 1943. A series of seven papers. *Am. J. Hyg.* 42: 1–105, 1945.

8. Francis, T., Jr., J. E. Salk and W. M. Brace. *J. A. M. A.* 131: 275, 1946.
9. Hirst, G. K., A. Vilches, O. Rogers and C. L. Robbins. *Am. J. Hyg.* 45: 96, 1947.
10. Francis, T., Jr., J. E. Salk and J. J. Quilligan, Jr. *Am. J. Pub. Health* 37: 1013, 1947.
11. Meiklejohn, G., D. L. Weiss, R. I. Shragg and E. H. Lennette. *Am. J. Hyg.* 55: 1, 1952.
12. Hilleman, M. R. *Proc. Soc. Exper. Biol. & Med.* 78: 208, 1951.
13. Magill, T. P. In press.
14. Meiklejohn, G., C. H. Kempe, W. G. Thalman and E. H. Lennette. *Am. J. Hyg.* 55: 12, 1952.
15. Archetti, I. and F. L. Horsfall, Jr. *J. Exper. Med.* 92: 441, 1950.
16. Salk, J. E., A. M. Laurent and M. L. Bailey. *Am. J. Pub. Health* 41: 669, 1951.
17. Salk, J. E., M. L. Bailey and A. M. Laurent. *Am. J. Hyg.* 55: 439, 1952.
18. Francis, T., Jr. *Cincinnati J. Med.* 31: 97, 1950.

FEDERATION PLACEMENT SERVICE

Information Regarding ***Annual*** *List of Fellowships*

DURING 1951 the Placement Service undertook the registration of fellowships available to individuals in the fields of experimental biology, and the preparation of a brief description of the essential facts and requirements necessary for application for each fellowship. In April of that year the first roster of such fellowships was issued and mailed to nearly 900 heads of departments and laboratories. A revised and more complete list describing approximately 90 fellowships, prepared six months later, was similarly circulated, and it was then decided to issue a new one annually in October. This date was selected as the latest time possible for the collection of definite information for the ensuing academic year which would still allow a reasonable interval for the dissemination of this information and for interested persons to file applications before the closing dates for the majority of the fellowships. Since these closing dates vary considerably, no one time of issuance can satisfactorily cover the whole field, but final dates for application usually fall in late winter or early spring, and October seems to provide the reasonable interval mentioned. In addition, organizations generally have information by that time of changes to be made affecting applications for the following year.

This list of fellowships has proved very popular, judging from the requests which have been received for it, and consequently seems to be meeting a distinct need. While it is recognized that many well-qualified students lack the means for continuing the research and training of their choice, it is equally true that many fellowships lack well-qualified applicants. In many cases this is due to the absence of information as to how and where to apply for this aid. If this list of fellowships can be made more nearly complete, with a copy available in all department and graduate school libraries, the necessary information will then be within the reach of most potential applicants.

When the 1951 register of fellowships was prepared, the term 'fellowship' was rather loosely applied. With the 1952 list, we shall conform to the National Research Council practice of restricting the term to gifts whose conditions of grant do not specify the rendering of services, and which are exempt from income tax. The teaching fellowships will be included in the same list but labeled 'teaching assistantships.' Of course it will be necessary to depend upon the information given us by those registering such opportunities, and it cannot be guaranteed that a listing falls into the tax-free class even though it has been called a 'fellowship.'

It will be difficult in some cases to separate the teaching assistantships from bona fide positions. To assist in this differentiation, it is suggested that teaching assistantships registered be available for short, definite periods of tenure, usually one year, and be open to people outside the awarding institution. In this way it is hoped to encourage the registration of the numerous university assistantships now available, so that these opportunities for work and study may become more widely known. Since the next list of fellowships, including the teaching assistantships, is to be prepared for distribution in October it will be necessary that information to be included reach the office before October 1, 1952.

During 1952 the list of fellowships is to be sent to the annual subscribers to the Placement Service lists. According to the schedule set up when charges were first made for the lists, all single copies have been priced at $5.00. It is hoped that, in the near future, the price for a single fellowship list can be lowered to $2.00. At first it may be possible to extend this lower price only to students, applicants and those ordering a large number of a single list. Some organizations have indicated an interest in securing multiple copies for use among their staff and students. Since it is desirable that all departments with graduate students have current information on fellowships, such a lower group rate should make it possible for a university to supply them even though all the departments do not find an annual subscription to all the lists useful.

Cards to register fellowships and teaching assistantships and order forms for this list may be secured by writing to the Placement Service, Federation of American Societies for Experimental Biology, 2101 Constitution Avenue Washington 25, D. C.

LETHA K. ANDERVONT, *Assistant to the Director, Federation Placement Service*